

# Repetitions in Strings

LUCIAN ILIE

Institut Gaspard-Monge, Université de Marne-la-Vallée  
Department of Computer Science, University of Western Ontario  
e-mail: `ilie@csd.uwo.ca`



# Repetitions

- **repetition** = substring with adjacent occurrences
- **square** – two occurrences
  - ATAGGATAGG = (ATAGG)<sup>2</sup>
  - hotshots = (hots)<sup>2</sup>
- **general (fractional) repetitions**
  - alfalfa = (alf) <sup>$\frac{7}{3}$</sup>
  - otavaltavaltavaltavaltiolta = o(taval) <sup>$\frac{21}{5}$</sup>  iolta
  - CGATATATATATATACGGC = CG(AT) <sup>$\frac{13}{2}$</sup>  CGGC
- for a repetition  $w^\alpha$ 
  - $\alpha$  is the **exponent**
  - the number of letters of  $w$  is the **period**



# Repetitions – Problems

- interesting **combinatorial problems**
  - count the number of various types of repetitions
  - squares already investigated by [Thue, 1906, 1912]
- useful **combinatorial algorithms**
  - find various types of repetitions
  - find **all** repetitions in **linear time**
- many **applications**
  - text algorithms
  - data compression
  - analysis of biological sequences



# Squares

- how many square occurrences in a string of length  $n$ ?
- all –  $\Theta(n^2)$  – AAAAAAAAAAAAAAAAAA
- primitively rooted square occurrences
  - ABABABAB – not primitively rooted
  - $\Theta(n \log n)$  – [Crochemore, 1981]
- squares (each square counted only once)
  - $2n$  – [Fraenkel, Simpson, 1998]
  - short proof – [Ilie, 2005]



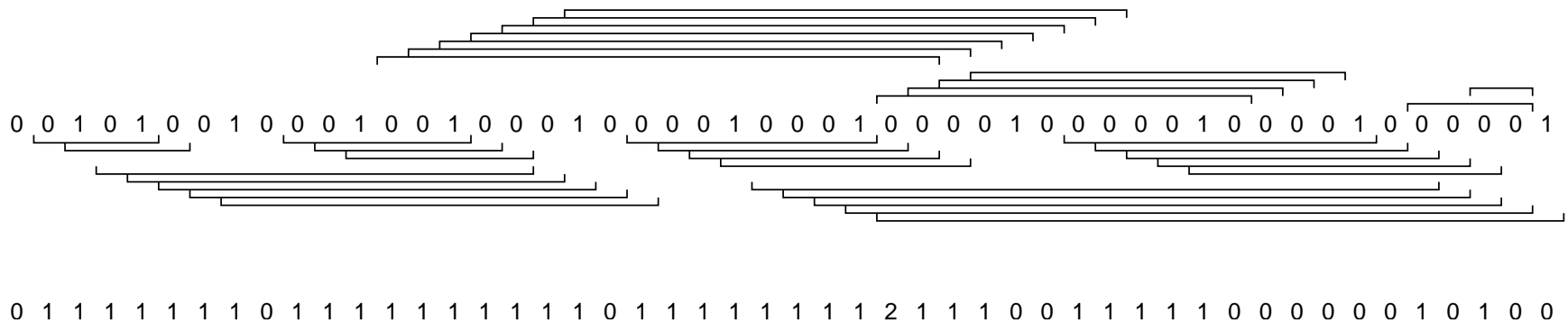
# Linearly many squares

- very important combinatorial result
- gives hope for identifying all repetitions in linear time
  - [Gusfield, Stoye, 2004]
- **conjecture**
  - [Fraenkel, Simpson, 1998], [Kolpakov, Kucherov, 2005]
  - $\leq n$  – supported by computations
  - probably very difficult to prove
    - it is optimal
    - examples – asymptotically  $n$  squares
- best bound –  $2n - \Theta(\log n)$  – [Ilie, 2005]



# Example with many squares

- $w_m = \bigodot_{i=1}^m 0^{i+1} 10^i 10^{i+1} 1$
- $|w_m| = \frac{3}{2}m^2 + \frac{13}{2}m$
- $\text{squares}(w_m) = \frac{3}{2}m^2 + 4m - 3 + \frac{\text{odd}(m)}{2}$
- $w_4$  and its squares are shown below



# All repetitions

- algorithms for finding all repetitions in linear time
  - problem – too many repetitions
  - solution – encode all more compactly
- [Crochemore, 1981, 1983] – linear for one square,  $\mathcal{O}(n \log n)$  for all primitively-rooted maximal integer
- [Apostolico, Preparata, 1983] –  $\mathcal{O}(n \log n)$  for all right-maximal
- [Main, Lorentz, 1985] –  $\mathcal{O}(n \log n)$  for all maximal
- [Main, 1989] – linear for all maximal leftmost
- [Iliopoulos, Moore, Smyth, 1997] – linear for all in Fibonacci



# Runs

- **run** – repetition that is
  - fractional
  - non-extendable (maximal)
  - primitively-rooted
- $o\underline{taval}taval\underline{taval}taval\underline{taval}iota = o(taval)^{\frac{21}{5}}iota$
- $CG\underline{ATATATATATATA}CGGC = CG(AT)^{\frac{13}{2}}CGGC$
- $00\underline{0101}000010 = 00(01)^{\frac{5}{2}}00010$
- $\underline{000}101\underline{0000}10 = 0^31010^410$



# Runs – linear bound

- all repetitions are encoded in runs
- number of runs in a string of length  $n$ 
  - $\mathcal{O}(n)$  – [Kolpakov, Kucherov, 1998]
  - compute all repetitions in linear time
    - modified Main's algorithm
    - based on Crochemore factorization
    - most important breakthrough – the bound
  - no constant could be derived from the proof
  - $5n$  – [Rytter, 2006]



# Runs – simple proof

- both proof are very complicated
  - Kolpakov, Kucherov – 8 (large) pages, contains case 2.1.2.2
  - Rytter – 9 pages, considers highly and weakly periodic runs
- [\[Crochemore, Ilie, 2006\]](#) – simple proof
  - 1.5 pages
  - improved Rytter's idea of “neighbour runs”



# Runs – sum of exponents

- $\mathcal{O}(n)$  – [Kolpakov, Kucherov, 1999]
  - sum of exponents – applications to analysis of algorithms
  - proof very complicated – 9 (large) pages, contains case 2.1.2.3.2
- [Crochemore, Ilie, 2006] – simple proof – 0.5 pages



# Number of runs – improved bounds

- analysis of any algorithm computing all repetitions
- **conjectures**
  - [Kolpakov, Kucherov, 1998] –  $\leq n$
  - [Franek, Simpson, Smyth, 2003] –  $\leq \frac{3}{2\phi}n = 0.927..n$
  - they proved this lower bound
- **bounds**
  - [Puglisi, Simpson, Smyth, 2006] –  $3.48n$
  - [Rytter, 2006] –  $3.44n$
  - [Crochemore, Ilie, 2006] –  $1.6n$ 
    - could be improved by computer verification to  $1.18n$  or lower



# Sum of exponents – improved bounds

- conjecture
- [Kolpakov, Kucherov, 1999] –  $\leq 2n$
- [Crochemore, Ilie, 2006] –  $5.6n$ 
  - could be improved by computer verification to  $2.9n$  or lower
  - the first explicit bound for the sum of exponents
  - from Rytter's paper –  $25n$  – “unsatisfactory”



# Proof ideas

- [Rytter, 2006]
  - runs counted at the start – logarithmically many
  - short period runs
  - long period runs
    - weakly periodic
    - highly periodic

abaababaabaababaabaabaabaababaaba



# Proof ideas

- [Rytter, 2006]
  - runs counted at the start – logarithmically many
  - short period runs
  - long period runs
    - weakly periodic
    - highly periodic

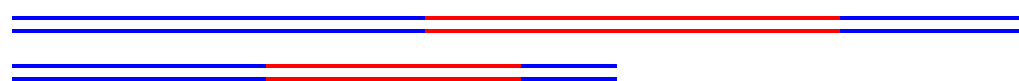
abaababaabaababaabaababaabaa



# Proof ideas

- [Rytter, 2006]
  - runs counted at the start – logarithmically many
  - short period runs
  - long period runs
    - weakly periodic
    - highly periodic

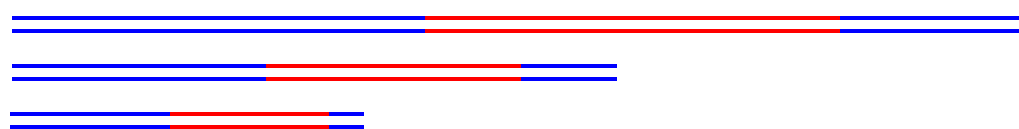
abaababaabaababaabaabaabaababaaba



# Proof ideas

- [Rytter, 2006]
  - runs counted at the start – logarithmically many
  - short period runs
  - long period runs
    - weakly periodic
    - highly periodic

abaababaabaababaabaabaabaabaaba



# Proof ideas

- [Rytter, 2006]
  - runs counted at the start – logarithmically many
  - short period runs
  - long period runs
    - weakly periodic
    - highly periodic

abaababaabaababaabaabaabaabaaba



# Proof ideas (2)

- [Crochemore, Ilie, 2006]
  - runs counted at the center – **linearly many**
  - short period runs
  - long period runs – 10 times better bound
  - for periods  $\geq 87$ 
    - we have  $0.06897n$  compared to  $0.67n$

abaabaabaabaabaababaabaabaabaabaaba



# Proof ideas (2)

- [Crochemore, Ilie, 2006]
  - runs counted at the center – **linearly many**
  - short period runs
  - long period runs – 10 times better bound
  - for periods  $\geq 87$ 
    - we have  $0.06897n$  compared to  $0.67n$

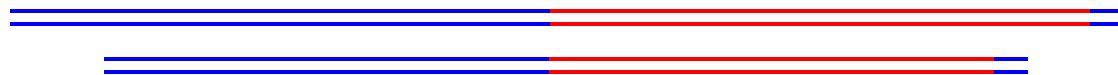
abaabaabaabaabaabaabaabaabaabaaba



# Proof ideas (2)

- [Crochemore, Ilie, 2006]
  - runs counted at the center – **linearly many**
  - short period runs
  - long period runs – 10 times better bound
  - for periods  $\geq 87$ 
    - we have  $0.06897n$  compared to  $0.67n$

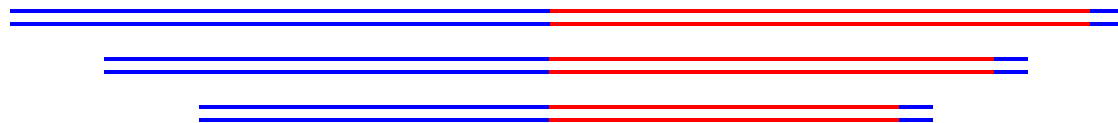
abaabaabaabaabaababaabaabaabaabaaba



# Proof ideas (2)

- [Crochemore, Ilie, 2006]
  - runs counted at the center – linearly many
  - short period runs
  - long period runs – 10 times better bound
  - for periods  $\geq 87$ 
    - we have  $0.06897n$  compared to  $0.67n$

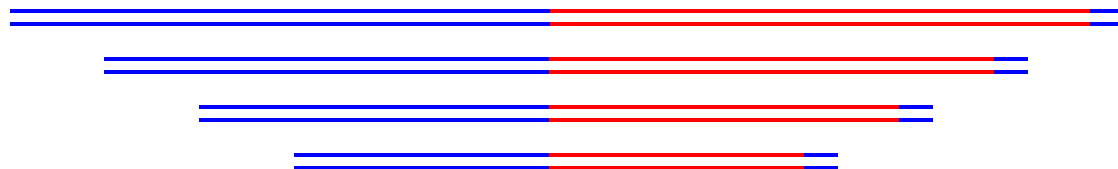
abaabaabaabaabaababaabaabaabaaba



# Proof ideas (2)

- [Crochemore, Ilie, 2006]
  - runs counted at the center – **linearly many**
  - short period runs
  - long period runs – 10 times better bound
  - for periods  $\geq 87$ 
    - we have  $0.06897n$  compared to  $0.67n$

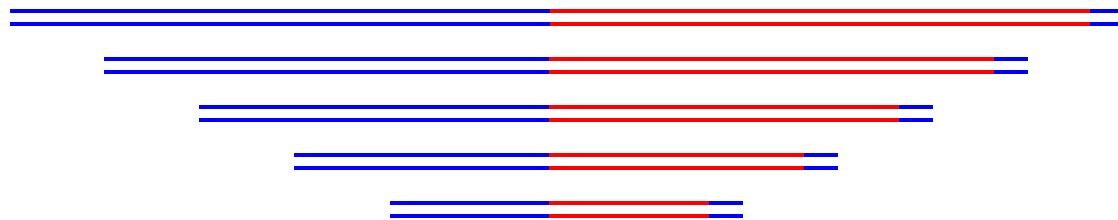
abaabaabaabaabaabaabaabaabaabaaba



# Proof ideas (2)

- [Crochemore, Ilie, 2006]
  - runs counted at the center – **linearly many**
  - short period runs
  - long period runs – 10 times better bound
  - for periods  $\geq 87$ 
    - we have  $0.06897n$  compared to  $0.67n$

abaabaabaabaabaabaabaabaabaabaaba





# Proof ideas (2)

- [Crochemore, Ilie, 2006]
  - runs counted at the center – linearly many
  - short period runs
  - long period runs – 10 times better bound
  - for periods  $\geq 87$ 
    - we have  $0.06897n$  compared to  $0.67n$

