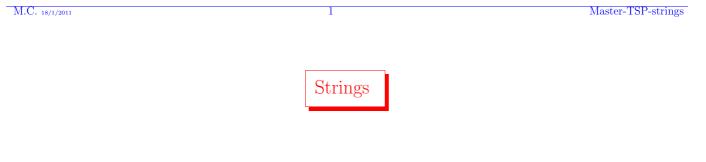
Text Searching and Processing

MAXIME CROCHEMORE

King's College London

Maxime.Crochemore@kcl.ac.uk http://www.dcs.kcl.ac.uk/staff/mac/



- * Alphabet: A (finite) set of letters, $A = \{a, b, c, ...\}$
- ***** Strings: A^* set of finite sequences of letters (ε denotes the empty string)
- * Length of a string x: |x| =length of the sequence
- * Notation—array representation: $x = x[0]x[1] \dots x[|x| 1]$

i	0	1	2	3	4	5	6	7	8
x[i]	b	a	b	a	a	b	a	b	a

- ★ Alphabet of a string: alph(x) set of letters occurring effectively in x; each letter of alph(x) appears at least once in x
- ***** Equality

$$x = y$$
 iff $|x| = |y|$ and $x[i] = y[i]$ for $i = 0, 1, ..., |x| - 1$

Factors

- **\star** Concatenation or product: xy is sequence x followed by sequence y
- ★ Factor: x factor of or occurs in y if y is a product uxv for two strings u, vx prefix of y if y = xv; x suffix of y if y = ux

i	0	1	2	3	4	5	6	7	8	
$\overline{y[i]}$	b	a	b	a	a	b	a	b	a	
left positions of aba		1			4		6			
right positions of aba				3			6		8	

- ★ **Positions**: x occurs in y at (left) position i if y = uxv and |u| = iequivalently $x = y[i]y[i+1] \dots y[i+|x|-1] = y[i \dots i+|x|-1]$
- ***** Positions of the first occurrence:

$$pos(x) = \min\{|u| : uxA^* \cap yA^* \neq \emptyset\}$$

★ Subsequence: x subsequence of y if $y = w_0 x[0] w_1 x[1] \dots x[|x| - 1] w_{|x|}$ for |x| + 1 strings $w_0, w_1, \dots, w_{|x|}$ equivalently, x can be obtained from y by deleting |y| - |x| letters

```
M.C. 18/1/2011
```

3

Master-TSP-strings

Powers

★ **Power**: u^k is the *k*th power of *u*, defined by $u^0 = \varepsilon$ and $u^e = u^{e-1}u$ for e = 1, 2, ..., k

Lemma 1

If $x^m = y^n$ for integers m, n > 0, then x, y are powers of the same string.

★ Primitive string: a (nonempty) string x is primitive if it is not the power of another string — equivalently x = u^k implies k = 1, and then x = u
abaab is primitive, while ε and bababa = (ba)³ are not

Lemma 2 (Primitivity Lemma)

x is primitive iff it is a factor of x^2 only as a prefix and as a suffix, that is, ux prefix of x^2 implies $u = \varepsilon$ or u = x

abaab occurs at positions 0, 5 only in abaababaab = $(abaab)^2$ bababa occurs at positions 0, 2, 4, 6 in babababababa = $(bababa)^2$ Proofs as exercises — consequences of the Periodicity Lemma * **Root of** x: unique primitive u for which $x = u^k$

Proposition 3

There exists one and only one primitive string which $x \neq \varepsilon$ is a power of.

abaab root of itself ba root of bababa

***** Conjugates: x, y are conjugates if x = uv and y = vu

abaab has 5 = |abaab| conjugates: abaab, baaba, aabab, ababa, babaa bababa has 2 = |ba| conjugates: bababa, ababab

Proposition 4

x, y are conjugate if and only if their roots are conjugate.

Proposition 5

x, y are conjugate if and only if there exists a string z such that xz = zy.

M.C. 18/1/2011

5

Master-TSP-strings