



# Chapter 32: String Matching

Arne Andersson

Informationsteknologi

UPPSALA UNIVERSITET

## Problem

Given a text of length  $n$  and a pattern of length  $m$ , where does the pattern occur in the text?

TATATCATATGCATATCATATATCATGAG

ATATCATG

Institutionen för informationsteknologi | www.it.uu.se

Informationsteknologi

UPPSALA UNIVERSITET

## Naive Algorithm

For each of the  $n$  positions in the text, try to fit the pattern at that position

Complexity  $O(nm)$

Institutionen för informationsteknologi | www.it.uu.se

Informationsteknologi

UPPSALA UNIVERSITET

## Rabin-Karp Algorithm

- Basic idea

Assume characters are digits

Example text:  
56232343567837837843234567654322

Pattern:  
78378

Institutionen för informationsteknologi | www.it.uu.se

Informationsteknologi

UPPSALA UNIVERSITET

561232343567837837843234567654322

56232

62323 = (56232-10000\*5) \*10+3

23234 = (62323-10000\*6) \*10+4

32343 = (23234-10000\*2) \*10+3

23435 ...etc...

34356

43567

35678

56783

67837

78378

83783

37837

78378

83784

Institutionen för informationsteknologi | www.it.uu.se

Informationsteknologi

UPPSALA UNIVERSITET

## Modular arithmetic

- $x \equiv y \pmod{q}$  means that  $x \bmod q = y \bmod q$  ( $x$  and  $y$  are equivalent mod  $q$ )
- Modular arithmetic: If  $a \equiv b$  and  $x \equiv y$ , then  $a+x \equiv b+y$  and  $ax \equiv by$
- Example: We calculate mod 17. Then  $19 \equiv 2$  and  $-3 \equiv 14$ , and  $19 - 3 \equiv 14 + 2 \pmod{17}$  ( $\equiv 16$ ) and  $19*(-3) \equiv 14*2 \pmod{17}$  ( $\equiv 11$ )

Institutionen för informationsteknologi | www.it.uu.se



### Use modular arithmetic to search for a fingerprint

561232343567837837843234567654322

56232

$$62323 = (56232 - 10000 * 5) * 10 + 3$$

$$23234 = (62323 - 10000 * 6) * 10 + 4$$

$$32343 = (23234 - 10000 * 2) * 10 + 3$$

23435 ...etc...

34356

43567

35678

56783

67837

78378

83783

37837

78378

83784

78378 mod 17 = 8, so if we do all calculations mod 17 we search for fingerprint 8

In order to do the arithmetic, we need to know 10000 mod 17, which is 4.



### Use modular arithmetic to search for a fingerprint

561232343567837837843234567654322

56232 13

$$62323 \quad 1 = ((13 - 4 * 5) * 10 + 3) \text{ mod } 17$$

$$23234 \quad 12 = ((1 - 4 * 6) * 10 + 4) \text{ mod } 17$$

$$32343 \quad 9 = ((12 - 4 * 2) * 10 + 3) \text{ mod } 17$$

23435 ...etc...

34356

43567

35678

56783

67837

78378

83783

37837

78378

83784