# Chapter 32: String Matching

Pierre Flener

(version of 2014-10-17)

# Problem

Given a text of length *n* and a pattern of length *m*, at what starting positions (shifts) does the pattern occur in the text?

Example text:

TATATCATATGCATATCATATATCATGAG

Pattern:

ATATCATG

Information Technology

UPPSALA UNIVERSITET

# Naive Algorithm

For each of the $n-m+1$ possible shifts, check whether the pattern (of length $m$) occurs with that shift in the text.

Complexity: $O((n-m+1)m)$ time at worst.

UPPSALA UNIVERSITET

Information Technology

# Rabin-Karp Algorithm

Basic idea: Assume characters are digits. Hence strings are numbers, which can be compared for equality in constant time.

Example text:

5623234356783783784323456765432 2

Pattern:

78378

UPPSALA UNIVERSITET

Information Technology

```
562323435678378378432345676543 22
56232
 62323      = (56232-10000•5)•10+3
  23234     = (62323-10000•6)•10+4
   32343    = (23234-10000•2)•10+3
    23435   …
     34356
      43567
       35678
        56783
         67837
          78378
           83783
            37837
             78378
              83784
```

Information Technology

UPPSALA UNIVERSITET

# Modular Arithmetic

- The notation $x \equiv y \pmod{q}$ means that $x \bmod q = y \bmod q$.
  We say that $x$ and $y$ are equivalent modulo $q$.
- Modular arithmetic:
  If $a \equiv b$ and $x \equiv y$, then
  $a + x \equiv b + y$
  $a \bullet x \equiv b \bullet y$
- Example, where we calculate modulo 17:
  $19 \equiv 2$ and $-3 \equiv 14$, hence
  $19 + (-3) \equiv 2 + 14$ $(\equiv 16)$
  $19 \bullet (-3) \equiv 2 \bullet 14$ $(\equiv 11)$

# Use Modular Arithmetic to Search for a Fingerprint

```
56232343467837837843234567654322
56232
 62323      = (56232-10000•5)•10+3
  23234     = (62323-10000•6)•10+4
   32343    = (23234-10000•2)•10+3
    23434   = (32343-10000•3)•10+4
     34346  = (23434-10000•2)•10+6
      43467 …
       34678
        46783
         67837
          78378
           83783
            37837
             78378
              83784
```

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

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

56232 mod 17 = 13, so the fingerprint of the first five characters is 13.

# Use Modular Arithmetic to Search for a Fingerprint

```
5623234346783783784323456767654322
56232      13
 62323      1 = ((13-4•5)•10+3) mod 17
  23234    12 = (( 1-4•6)•10+4) mod 17
   32343    9 = ((12-4•2)•10+3) mod 17
    23434   …
     34346
      43467
       34678
        46783
         67837
          78378
           83783
            37837
             78378
              83784
```

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

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

56232 mod 17 = 13, so the fingerprint of the first five characters is 13.

# Use Modular Arithmetic to Search for a Fingerprint

```
5623234346783783784323456765432 2
56232        13
 62323        1 = ((13-4•5)•10+3) mod 17
  23234      12 = (( 1-4•6)•10+4) mod 17
   32343      9 = ((12-4•2)•10+3) mod 17
    23434     8 = (( 9-4•3)•10+4) mod 17
     34346
      43467
       34678
        46783
         67837
          78378
           83783
            37837
             78378
              83784
```

Spurious hit!
Fingerprint is 8,
But '23434' ≠ '78378'

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

Information Technology

UPPSALA UNIVERSITET

# Use Modular Arithmetic to Search for a Fingerprint

```
56232343467837837843234567654322
56232          13
 62323          1 = ((13-4•5)•10+3) mod 17
  23234        12 = (( 1-4•6)•10+4) mod 17
   32343        9 = ((12-4•2)•10+3) mod 17
    23434       8 = (( 9-4•3)•10+4) mod 17
     34346      6 = (( 8-4•2)•10+6) mod 17
      43467    15 = (( 6-4•3)•10+7) mod 17
       34678   15 = ((15-4•4)•10+8) mod 17
        46783   9 = ((15-4•3)•10+3) mod 17
         67837  7 = (( 9-4•4)•10+7) mod 17
          78378 8 = (( 9-4•6)•10+8) mod 17
           83783
            37837
             78378
              83784
```

Another hit.
Check '78378' = '78378'
Success!

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