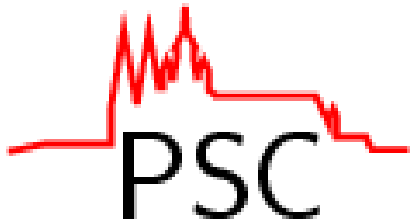
The String Matching Algorithm Research Tool

Simone Faro, Thierry Lecroq, Stefano Borzi, Simone Di Mauro
and Alessandro Maggio

Dipartimento di Matematica e Informatica, Università di Catania, Italy
LITIS, University of Rouen, France

Prague Stringology Conference
29 – 31 August 2016 – Prague, Czech Republic



Exact String Matching

Problem

Searching for all exact occurrences of a pattern x ($|x| = m$) in a text y ($|y| = n$)

Solutions (KMP, BM, ...)

Preprocessing of the pattern and use of a sliding window

SMART

String Matching Algorithms Research Tool

- implemented in the C programming language
- repository of implemented algorithms and text corpora
- framework to evaluate the performances of algorithms
- possibility of easily plug new algorithms

State of the Art

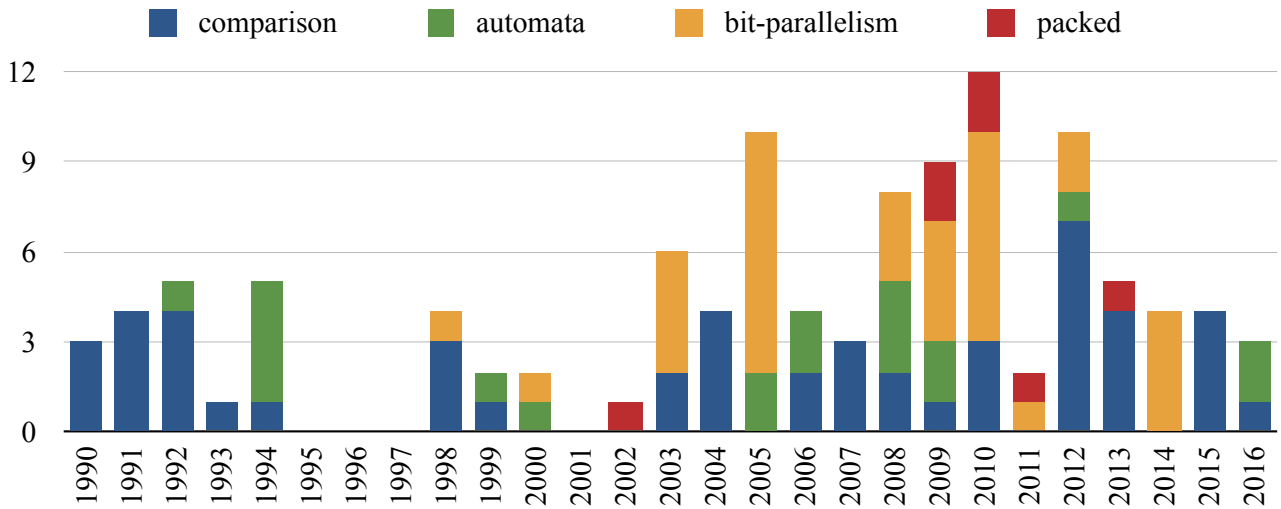
- many algorithms
- Hume and Sunday framework, 1991
- Charras and Lecroq ESMAJ, PSCW'98
- some surveys: Faro and Lecroq, PSC'11

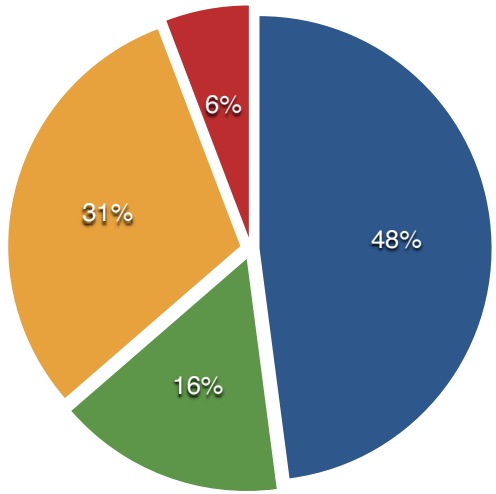
Implemented algorithms

120 algorithms (300 variants)

- classical
- most efficient
- hardware aware
- more in:
  S. Faro
  Exact online string matching bibliography
  CoRR, abs/1605.05067, 2016

Implemented algorithms





Text Corpora

- 2 English texts (6.1 MB, 94 characters)
- 7 Italian texts (5 MB, 120 characters)
- 7 French texts (6.6 MB, 119 characters)
- 5 Chinese texts (6.6 MB, 160 characters)
- *E. coli* (4.4 MB, 4 characters)
- a set of protein sequences (3.1 MB, 20 characters)
- a set of midi sequences (2.7 MB, 117 characters)
- `rand`σ random texts over an alphabet of size σ, uniform distribution, $\sigma \in \{2, 4, 8, 16, 32, 64, 128, 256\}$
- one can easily add a new corpus

Running experiments

One can easily:

- select/unselect algorithms
- launch smart with different parameters
- plug your own algorithms

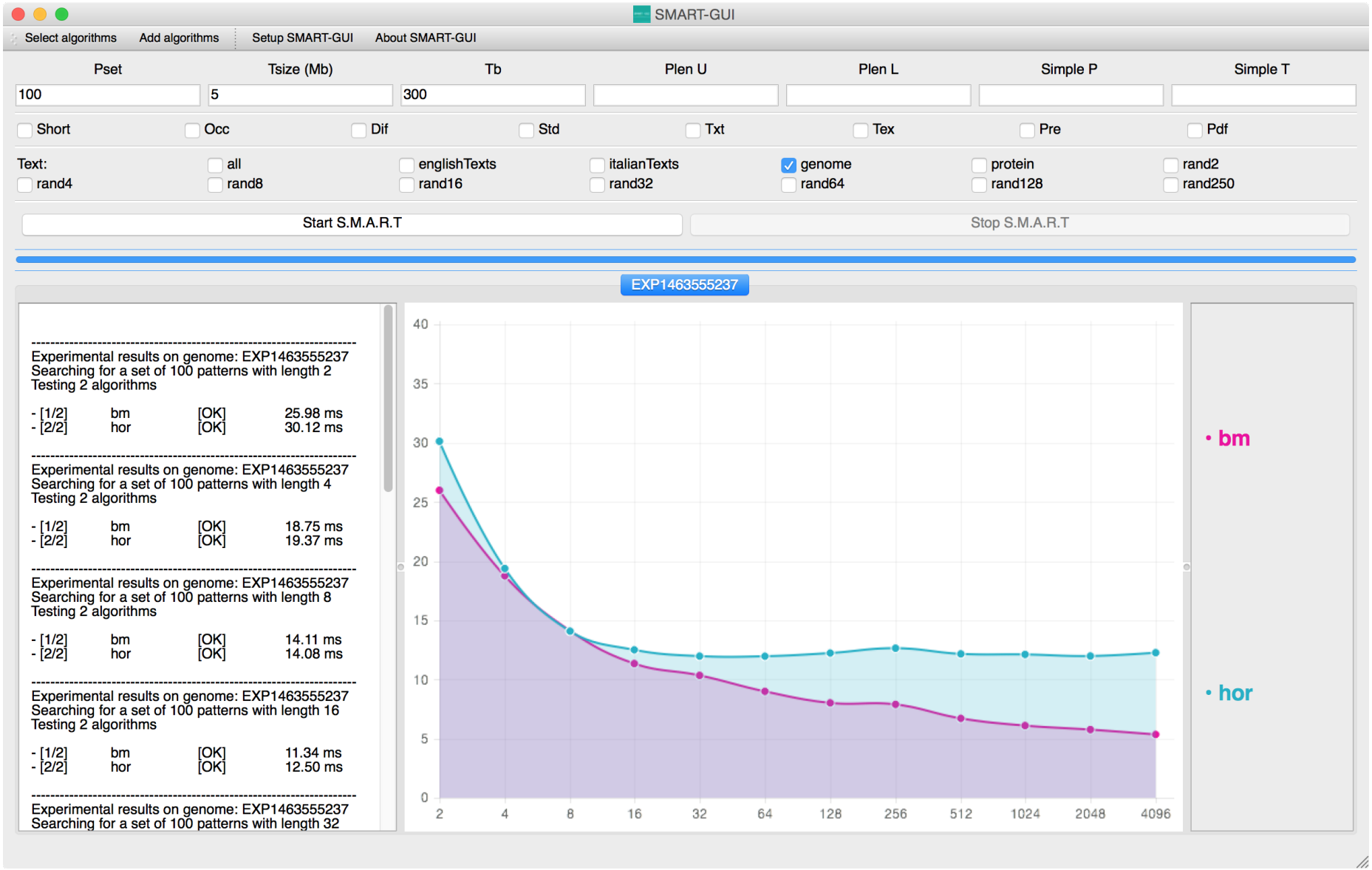
Outputs

- simple text
- LaTeX
- xml
- html
- php

GUI

- implemented in C++
- Qt WebKit

GUI


SMART-GUI
Select algorithms
Add algorithms
Setup SMART-GUI
About SMART-GUI
Pset
Tsize (Mb)
Tb
Plen U
Plen L
Simple P
Simple T
100
5
300
Short
Occ
Dif
Std
Txt
Tex
Pre
Pdf
Text:
all
englishTexts
italianTexts
genome
protein
rand2
rand4
rand8
rand16
rand32
rand64
rand128
rand250
Start S.M.A.R.T
Stop S.M.A.R.T
EXP1463555237
Experimental results on genome: EXP1463555237
Searching for a set of 100 patterns with length 2
Testing 2 algorithms
- [1/2] bm [OK] 25.98 ms
- [2/2] hor [OK] 30.12 ms
Experimental results on genome: EXP1463555237
Searching for a set of 100 patterns with length 4
Testing 2 algorithms
- [1/2] bm [OK] 18.75 ms
- [2/2] hor [OK] 19.37 ms
Experimental results on genome: EXP1463555237
Searching for a set of 100 patterns with length 8
Testing 2 algorithms
- [1/2] bm [OK] 14.11 ms
- [2/2] hor [OK] 14.08 ms
Experimental results on genome: EXP1463555237
Searching for a set of 100 patterns with length 16
Testing 2 algorithms
- [1/2] bm [OK] 11.34 ms
- [2/2] hor [OK] 12.50 ms
Experimental results on genome: EXP1463555237
Searching for a set of 100 patterns with length 32
40
35
30
25
20
15
10
5
0
2
4
8
16
32
64
128
256
512
1024
2048
4096
bm
hor

GUI

Report of Experimental Results
Test Code EXP1463554145
Date 2016:05:18 08:49:06
Text rand2 (alphabet : 2 - size : 5242880 bytes)

	2	4	8	16	32	64	128	256	512
BM	36.20	32.12	25.39	19.25	14.19	11.93	9.77	8.76	7.71
HOR	39.74	34.27	35.85	37.20	35.70	36.41	37.68	37.18	37.20

Table 1. Running times of experimental tests n.EXP1463554145. Each time value is the mean of 100 runs. Running times are in milliseconds.

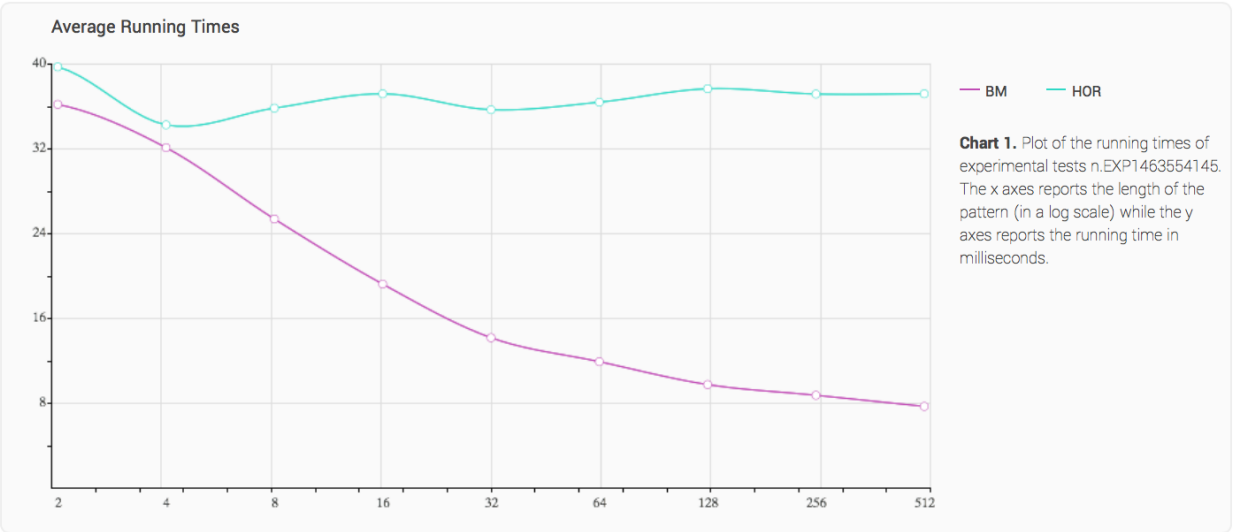
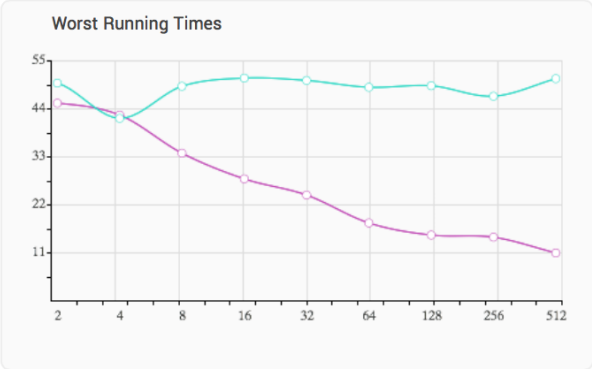
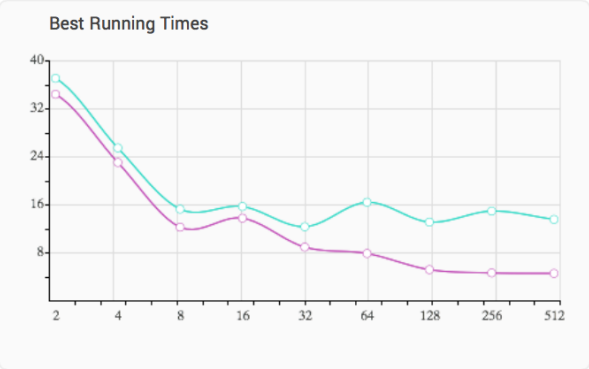


Chart 1. Plot of the running times of experimental tests n.EXP1463554145. The x axes reports the length of the pattern (in a log scale) while the y axes reports the running time in milliseconds.





Experiments

Efficiency

- mean of running times over a large set of runs
- includes preprocessing

Stability

standard deviation of running times

Flexibility

ability of an algorithm to perform well in different situations

Conclusion

- available at `http://www.dmi.unict.it/~faro/smart/` and `https://github.com/smart-tool`
- already used in different studies
- must take into account 128-bit machines

Thank you for your attention!