Boyer-Moore algorithm: Heuristics performance comparison

GENOME INFORMATICS(13M111GI)
UNIVERSITY OF BELGRADE, SCHOOL OF ELECTRICAL ENGINEERING
Dušan Damljanović 19/3461

Content

- Boyer-Moore algorithm
- Bad character rule
- Good suffix rule
- BMH
- Performance analysis

Boyer-Moore algorithm

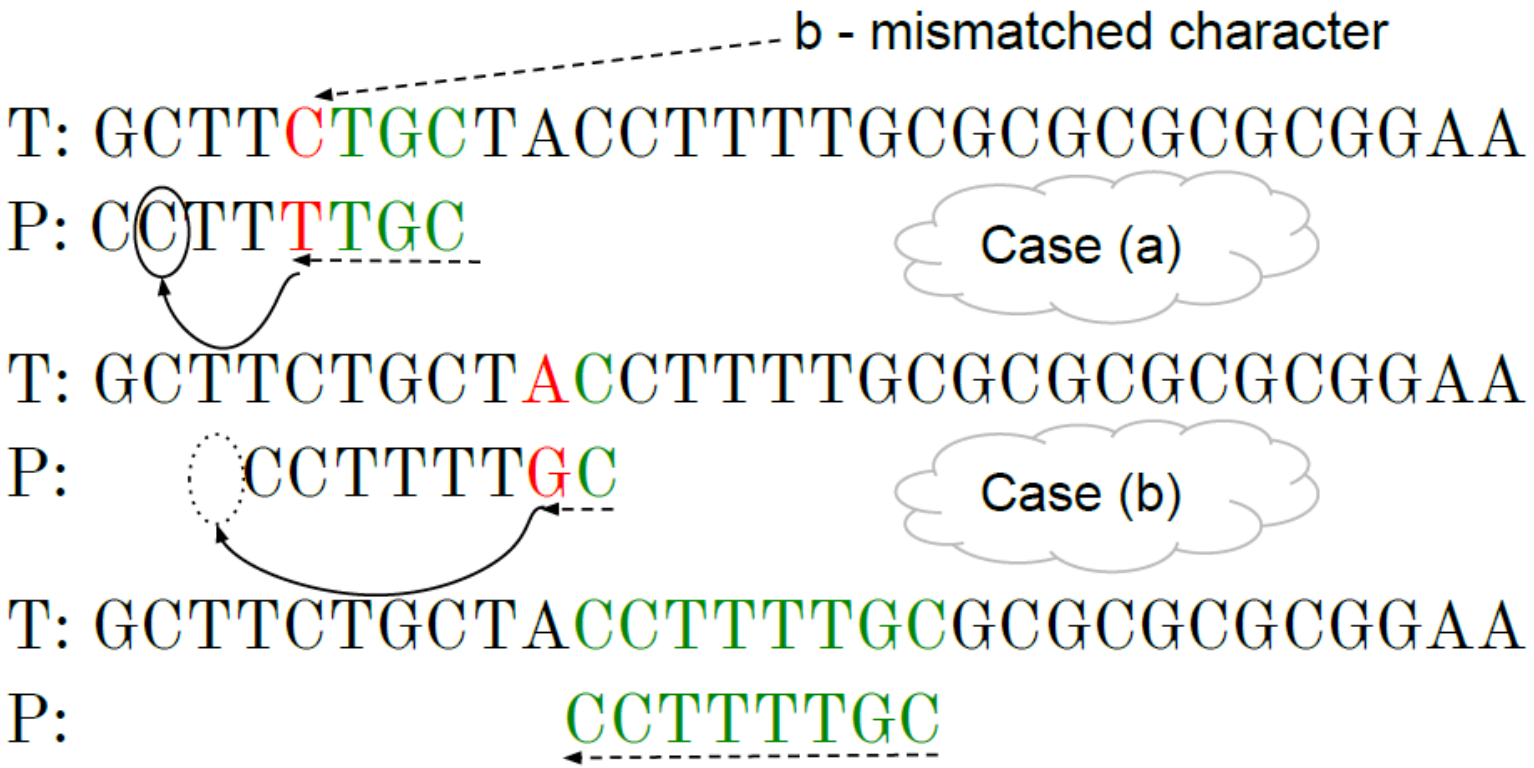
- efficient string-searching algorithm
- the standard benchmark for practical string-search literature
- online exact matching
- uses knowledge gained from character comparisons to skip future alignments

Boyer-Moore algorithm

- heuristic
- preprocesses
- match pattern and text
- mismatch → shift pattern
- matched → print and shift pattern

Bad character rule

- mismatch -> skip alignments



Good suffix rule

- t is the longest suffix of P that matches T in the current position
- shift P

BMH

- algorithm only uses the bad characters shift
- the size of shift is determined by the character in the text string which is aligned to the last one of pattern string

0 1 2 3
a b c d

Letter	a	b	c	d	*
Value	3	2	1	4	4

Value (a) = 4 - 0 - 1 = 3

Value (b) = 4 - 1 - 1 = 3

Value (c) = 4 - 2 - 1 = 3

Value (d) = 4

BMH

- easy to implement in case of mismatch
- the removal of Good-Suffix might decrease the shift

	S	T	R	I	N	G	M	A	T	C	H	I	N	G	I	S	T	O	F	I	N	D	T	H	E	P	A	T	T	E	R	N
1	P	A	T	T	E	R	N																										
2								P	A	T	T	E	R	N																			
3															P	A	T	T	E	R	N											
4																						P	A	T	T	E	R	N				
5																									P	A	T	T	E	R	N	
6																										P	A	T	T	E	R	N

Horspool sundays

- mismatch on the first compared character or on complete match:
- if (1st) not in the text: move len(pattern) + 1
- if (2nd) not in the text: move len(pattern) + 2
- if the move for the first letter is 1, do it
- else move for the max value of calculated moves for first and second letter

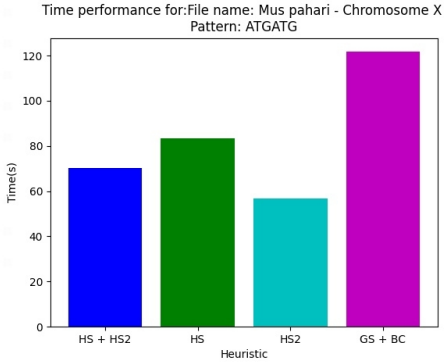
Horspool Sundays

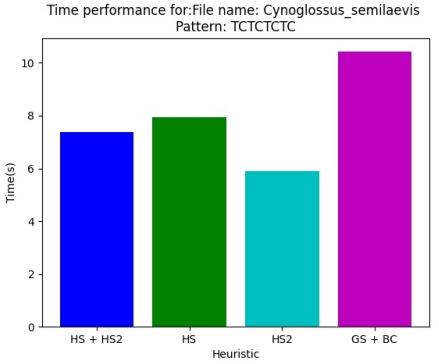
- mismatch on other than first character:
- if (first) not in text: move len(pattern) + 1
- if (second) not in text move len(pattern) + 2
- if move for the first letter is 1 and appearances(n) in unmatched pattern is 1:
- move for max value of calculated moves for the second letter
- len(pattern) + 1
- else move for the max value of calculated moves for the first and second letter

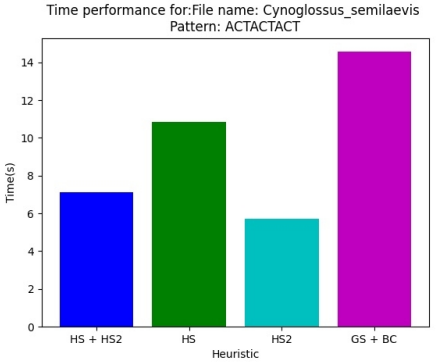
Performance analysis

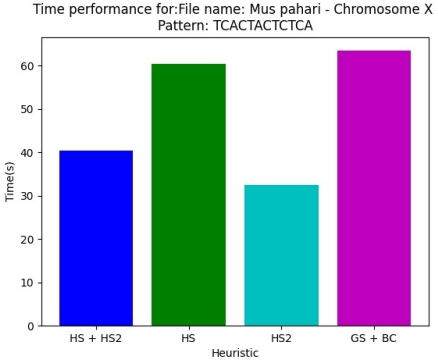
Pattern	File
ATGCATG	Coffea arabica - Chromosome 1c
TCTCTCTA	Coffea arabica - Chromosome 1c
TTCACTACTCTCA	Coffea arabica - Chromosome 1c
ATGATG	Mus pahari - Chromosome X
CTCTCTA	Mus pahari - Chromosome X
TCACTACTCTCA	Mus pahari - Chromosome X
ACTACTACT	Cynoglossus semilaevis – Chromosome 10
TCTCTCTC	Cynoglossus semilaevis – Chromosome 10
AAGTTAGAAA	Cynoglossus semilaevis – Chromosome 10

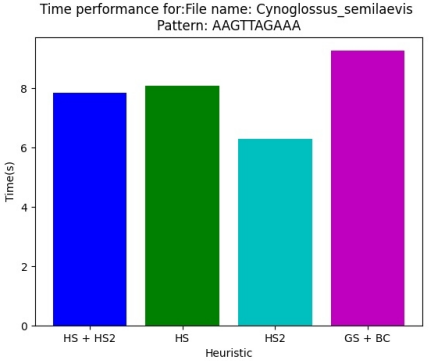
Time analysis

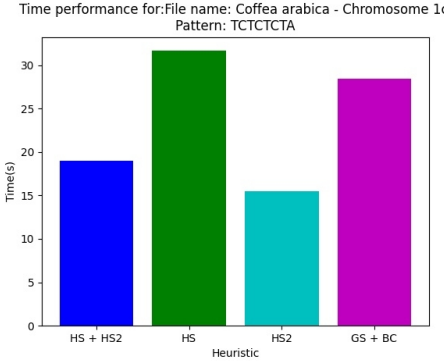


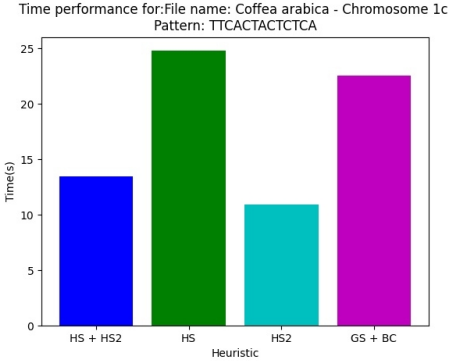


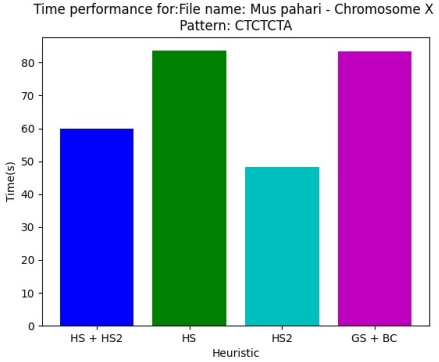


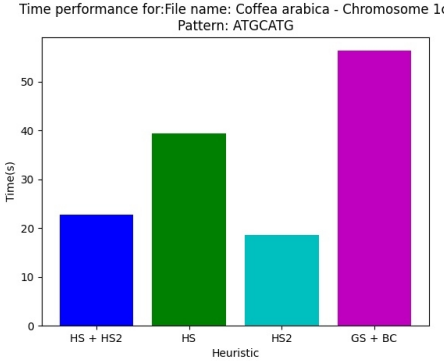








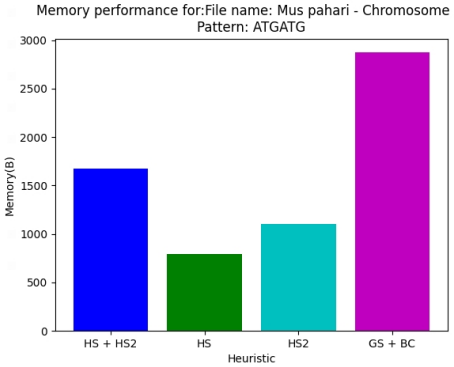


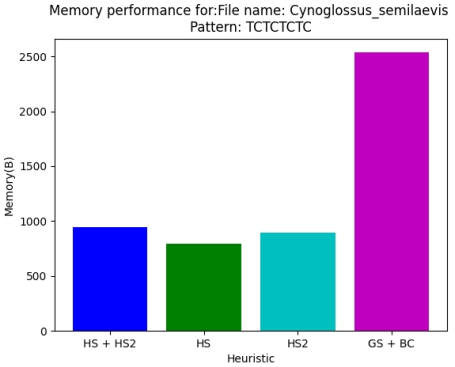


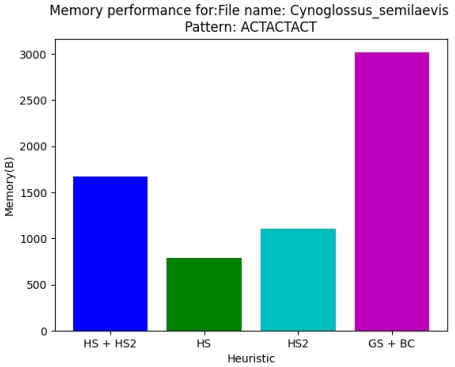
Time analysis

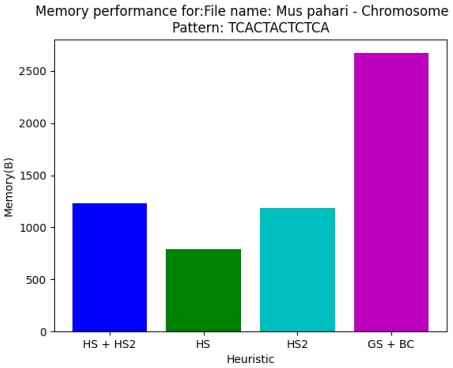
	HS + HS2	HS	HS2	GS + BC
File name: Mus pahari - Chromosome X: ATGATG	70.3	83.53	56.72	121.69
File name: Cynoglossus_semilaevis: TCTCTCTC	7.39	7.95	5.91	10.42
File name: Cynoglossus_semilaevis: ACTACTACT	7.13	10.83	5.7	14.56
File name: Mus pahari - Chromosome X: TCACTACTCTCA	40.33	60.37	32.52	63.43
File name: Cynoglossus_semilaevis: AAGTTAGAAA	7.85	8.08	6.3	9.26
File name: Coffea arabica - Chromosome 1c: TCTCTCTA	19.03	31.65	15.48	28.42
File name: Coffea arabica - Chromosome 1c: TTCACTACTCTCA	13.48	24.8	10.93	22.55
File name: Mus pahari - Chromosome X: CTCTCTA	59.94	83.6	48.25	83.47
File name: Coffea arabica - Chromosome 1c: ATGCATG	22.68	39.47	18.64	56.33

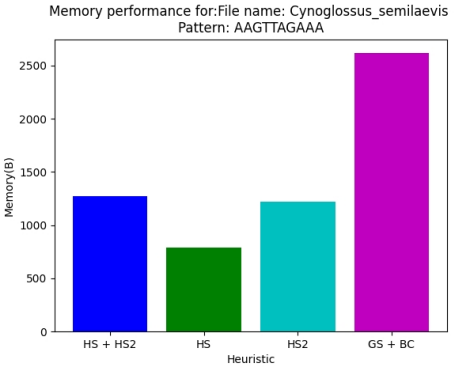
Memory analysis

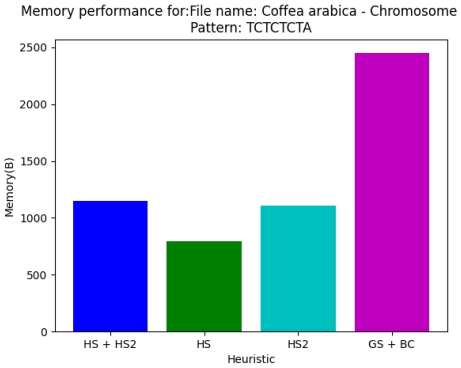


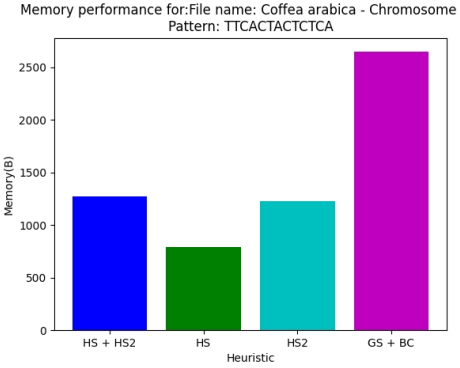


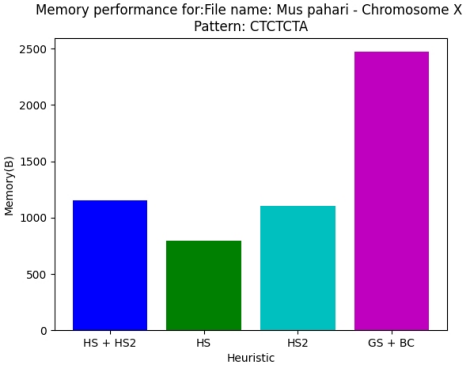


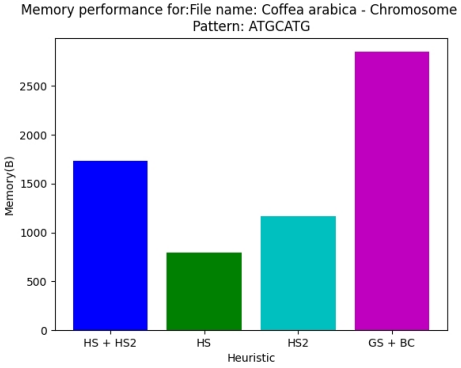












Memory analysis

	HS + HS2	HS	HS2	GS + BC
File name: Mus pahari - Chromosome X: ATGATG	1672	792	1104	2872
File name: Cynoglossus_semilaevis: TCTCTCTC	944	792	896	2536
File name: Cynoglossus_semilaevis: ACTACTACT	1672	792	1104	3016
File name: Mus pahari - Chromosome X: TCACTACTCTCA	1232	792	1184	2672
File name: Cynoglossus_semilaevis: AAGTTAGAAA	1272	792	1224	2616
File name: Coffea arabica - Chromosome 1c: TCTCTCTA	1152	792	1104	2448
File name: Coffea arabica - Chromosome 1c: TTCACTACTCTCA	1272	792	1224	2648
File name: Mus pahari - Chromosome X: CTCTCTA	1152	792	1104	2472
File name: Coffea arabica - Chromosome 1c: ATGCATG	1736	792	1168	2848