Ausarbeitung zur Implementierungsaufgabe 1 – AMBI 2021

Magdalena Weber und Martin Brand

21. Juli 2021

Test 1

Die Datei **text.fasta** enthält einen deutschen Text (im Fasta-Format). Suchen Sie in der Datei nach den Wörtern **Besen**, **Wasserstroeme**, **Eimer**.

Ergebnisse

Muster	Anzahl der Treffer	Shifts
Besen	5	[392, 1175, 1550, 2383, 2390]
Wasserstroeme	1	[1514]
Eimer	0	-

Laufzeiten und Suchschritte

Naive String Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
Besen	0.003980398178100586	10000
Wasserstroeme	0.0011065006256103516	25896
Eimer	0.0009975433349609375	10000

Rabin Karp Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
Besen	0.001989603042602539	1315
Wasserstroeme	0.0	2769
Eimer	0.002028942108154297	945

Knuth Morris Pratt Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
Besen	0.0019476413726806 64	2519
Wasserstroeme	0.003190279006958008	2537
Eimer	0.0020003318786621094	2516

Boyer Moore Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
Besen	0.0009980201721191406	110
Wasserstroeme	0.0	48
Eimer	0.0005249977111816406	77

Test 2

Die Datei **Virus.fasta** enthält das Genom eines Virus. Suchen Sie in diesem nach den Sequenzen GTATTA, TTTCGAAA, AAATTGACG.

Ergebnisse

Muster	Anzahl der Treffer	Shifts
GTATTA	32	[0, 27, 54, 81, 108, 135, 664, 910, 2302, 5451, ...]
TTTCGAAA	3	[574, 26729, 40032]
AAATTGACG	0	-

Laufzeiten und Suchschritte

Naive String Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
GTATTA	0.008044242858886719	240252
TTTCGAAA	0.008076906204223633	320320
AAATTGACG	0.011819839477539062	360351

Rabin Karp Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
GTATTA	0.030391216278076172	21936
TTTCGAAA	0.027830123901367188	29216
AAATTGACG	0.028927087783813477	32229

Knuth Morris Pratt Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
GTATTA	0.038549184799194336	46329
TTTCGAAA	0.05999612808227539	52925
AAATTGACG	0.03966546058654785	52058

Boyer Moore Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
GTATTA	0.030913829803466797	7404
TTTCGAAA	0.022725582122802734	5149
AAATTGACG	0.02213287353515625	3244

Test 3

Die Datei **BA000002.fna** enthält das Genom eines Archaebakterium. Suchen Sie darin nach den Restriktionsschnittstellen GAATTC, GGATCC und ATTTAAAT von EcoRI, BamHI und SwaI.

Ergebnisse

Muster	Anzahl der Treffer	Shifts
GAATTC	126	[2743, 7589, 37479, 62296, 65809, 67859, 69679, ...]
GGATCC	397	[1268, 8198, 8476, 9395, 13755, 14596, 17703, ...]
ATTTAAAT	15	[513279, 639142, 736136, 737696, 737702, 739889, ...]

Laufzeiten und Suchschritte

Naive String Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
GAATTC	0.36690449714660645	10018140
GGATCC	0.36446452140808105	10018140
ATTTAAAT	0.4149892330169678	13357504

Rabin Karp Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
GAATTC	0.8705151081085205	926820
GGATCC	0.9073717594146729	911394
ATTTAAAT	0.9503421783447266	1213768

Knuth Morris Pratt Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
GAATTC	1.5476446151733398	2136418
GGATCC	1.6134674549102783	2103389
ATTTAAAT	1.483191728591919	2029480

Boyer Moore Matcher

Muster	Laufzeit in Sekunden	Anzahl der Suchschritte
GAATTC	0.7921574115753174	187629
GGATCC	0.8875079154968262	237024
ATTTAAAT	0.4982485771179199	98639

Test 4

Suchen Sie außerdem im Archaebakterium Genom nach dem Gen, welches in der Datei **gen.fasta** gegeben ist.

Es gibt einen Treffer mit dem Shift 92380.

Algorithmus	Laufzeit	Anzahl der Suchschritte
Naive String	0.5716068744659424	2452292220
Rabin Karp	0.9951279163360596	223291530
Knuth Morris Pratt	1.5147531032562256	2125010
Boyer Moore	0.1896343231201172	29240