Übung 2
Algorithmische Bioinformatik

WS 15/16

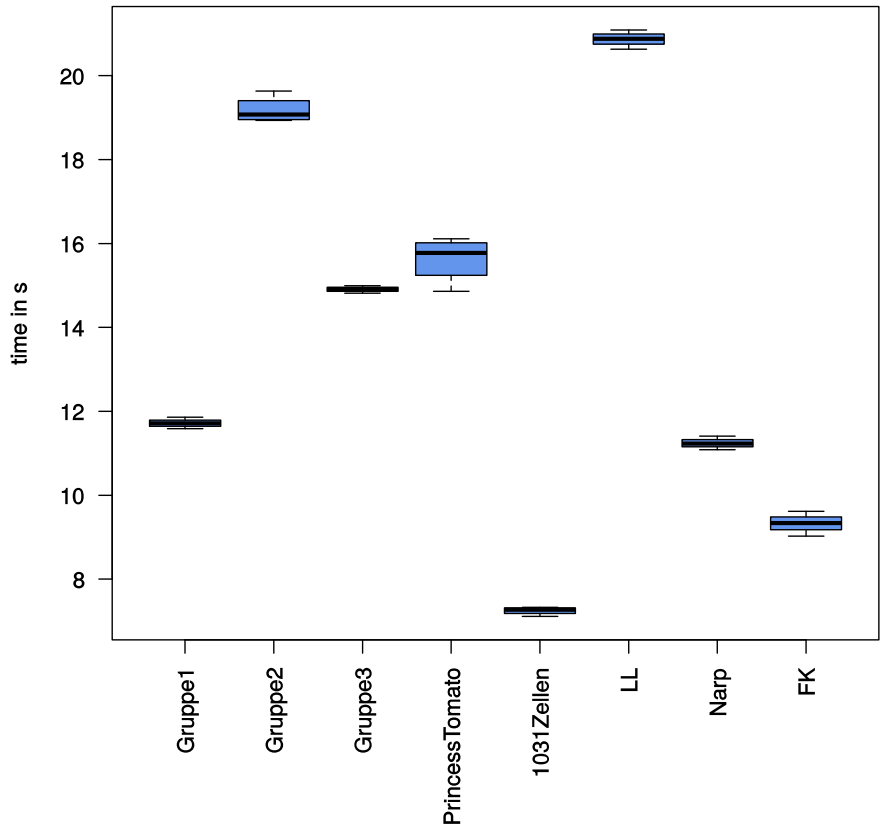
Yvonne Mayer

Lösungen/ Wettbewerb Übung 1

Lösungen Übung 1 vorstellen

- Fasta Sequenzen einlesen (5P)
- Implementierung String-Matching Algorithmus (8P)
- Abhängigkeit Patternlänge von Laufzeit (7P)
- (Einfluss Alphabetgröße)

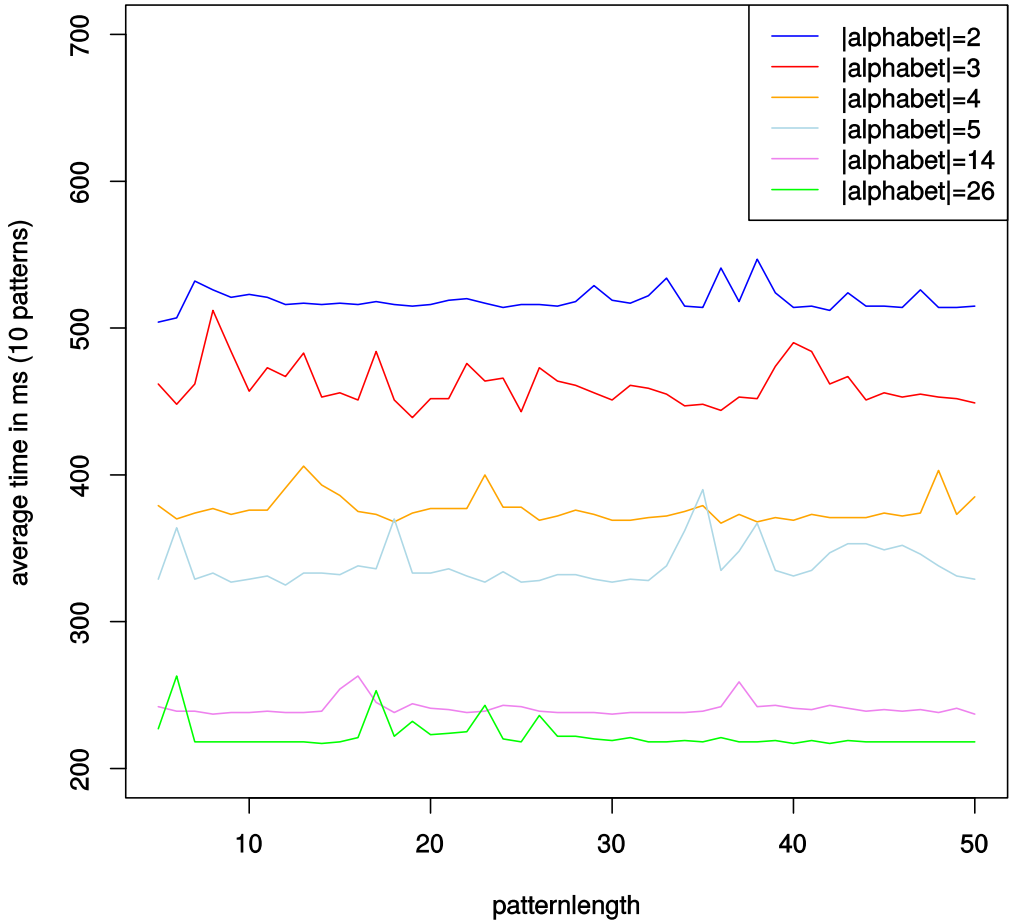
Wettbewerb


time in s
20
18
16
14
12
10
8
Gruppe1
Gruppe2
Gruppe3
PrincessTomato
1031Zellen
LL
Narp
FK

Wettbewerb

	Gruppe1	Gruppe2	Gruppe3	Princess Tomato	1031 Zeilen	LL	Narp	FK
Challenge1	0	0	0	0	3	0	1	2

Naive String Matching: Alphabet Size and Pattern Length



- random template: 500000 letters
- 10 random patterns for each length

Übung 2

Übung 2.1

(1) Boyer-Moore (10 Punkte)

- Implementieren Sie den Boyer-Moore Algorithmus. Hierfür müssen sowohl „Bad-Character Rule“ (BCR) als auch „Good Suffix Rule“ (GSR) implementiert und eingesetzt werden. (6P)

- Erstellen Sie einen Plot, der die Laufzeit in Abhängigkeit der Patternlänge (5-50bp) darstellt. Beschreiben Sie kurz das Ergebnis. (3P)

- Für den Wettbewerb (Parameter --challenge) ist die Verwendung von BCR und GSR optional. Natürlich dürfen auch vorgestellte Verbesserungen der GSR/BCR (bzw. Erweiterung mit linearem Worst Case) implementiert werden. Begründen Sie ihre Wahl für den Wettbewerb. (1P)

Übung 2.2

(2) Erwartungswert (3 Punkte)

- Geben Sie für jedes Pattern zusätzlich an, wieviele Fundstellen Sie in dem Template per Zufall erwartet hätten.
- Berücksichtigen Sie hierfür die beobachteten Häufigkeiten der einzelnen Basen im Template.

Übung 2.3

(3) Verständnisfragen zu Boyer-Moore (4 Punkte)

- Wie kann das Pattern-Preprocessing in linearer Zeit und linearem Platz durchgeführt werden? (3 Punkte)
- Welche Eigenschaft sollten Alphabete besitzen um von der Good Suffix Rule zu profitieren? Begründen Sie Ihre Antwort (1 Punkt)

Übung 2.4

(4) Anzahl der Fundstellen (3 Punkte)

- Im Testdatensatz ist eine Folge von mehreren Pattern enthalten, die jeweils nur aus „a“ bestehen und unterschiedliche Länge haben.
- Erklären Sie, warum die Häufigkeiten der Vorkommen der längeren Pattern so langsam abnehmen.

Programmaufruf

- Auf der Übungs-Homepage sind zwei Dateien bereitgestellt, die das Template und mehrere Patterns enthalten
  - Template: Chromosom 20 des Menschen (50MB)
  - Patterns: Schnittstellen von Restriktionsenzymen

- Programm muss wie folgt aufrufbar sein
  - für die Boyer-Moore Implementierung:
    ```
    java -jar Assignment1_GrXY.jar file1 file2
    ```
  - für den Wettbewerb:
    ```
    java -jar Assignment1_GrXY.jar file1 file2 --challenge
    ```
  - file1: Dateiname Patterns, file2: Dateiname Template
  - file1 und file2 sind unkomprimiert

Programmausgabe

- Ausgabe auf STDOUT
- Pro Paar (Template/Pattern) muss das Programm dann ausgeben (in jeweils neuer Zeile)
  - Pattern
  - Patternlänge
  - Anzahl Fundstellen
  - Anzahl erwarteter Fundstellen
  - Anzahl ausgeführter Zeichenvergleiche
  - Startpositionen der ersten zehn Fundstellen

```
> tccgga
> Length: 6
> Occurrences: 2506
> Expected: 1234
> Char-Comparisons: 1234567
> Positions: 29561, 30666, 134809, 244141, 276753, 315061, 318465, 330539, 344994, 347335
```

Wettbewerb

- Umfasst Aufgabe 2.1: Entscheidet wie der Algorithmus für die Challenge (GCR/BCR?) verwendet werden soll (Parameter --challenge)

- Wir messen die Gesamtlaufzeit Ihres Verfahrens über alle Pattern (mittels Linux „`time`“)

- Wur verwenden 20 neue Patterns, deren Länge zwischen 5 und 64 Zeichen liegt (Alphabet E = {ACGTN})

- Wettbewerbspunkte:
  - Platz 1: 3 Punkte
  - Platz 2: 2 Punkte
  - Platz 3: 1 Punkt

Abgabe

- Abgabe bis Sonntag den 15.11.2015 um 23:59 Uhr
- Abgabe per Email an: mayeryvo@informatik.hu-berlin.de
  (gerne auch Fragen zur Übung per Email)
  - PDF mit Plot (2.1) und Antworten zu allen Fragen
    (2.1: Begründung Wettbewerb, 2.2: kurze Erläuterung, 2.3, 2.4)
  - Jar Datei mit Quellcode (Dateiname: AssignmentX_GrXY.jar,
    jar ggf. als rar verpacken)
  - Kompilierung unter Java 1.7, Jar Datei auf gruenau2 testen,
    Abgaben ohne Quellcode werden ignoriert!
- Bitte die geschätze Bearbeitungszeit mitteilen
  (z.B. 10h für zwei Teilnehmer)

Zur Orientierung

Anzahl Vorkommen der Pattern im Template

- tccgga: 2506
- gctacc: 6799
- taataa: 28279
- cctcagc: 17520
- cctgcagg: 2425
- ggcgcgcc: 141
- ccccccccccc: 140
- aaaaaaaaaaa: 52695
- aaaaaaaaaaaa: 44140
- aaaaaaaaaaaaaaa: 25063
- aaaaaaaaaaaaaaaaaaaa: 8571

Optional

Boyer-Moore

- Welche der beiden Regeln wird bei den vorgegebenen Pattern häufiger verwendet? Wie hoch ist die durchschnittliche Sprunglänge für GSR und BCR?