Übung 6
Algorithmische Bioinformatik

WS 15/16

Yvonne Mayer

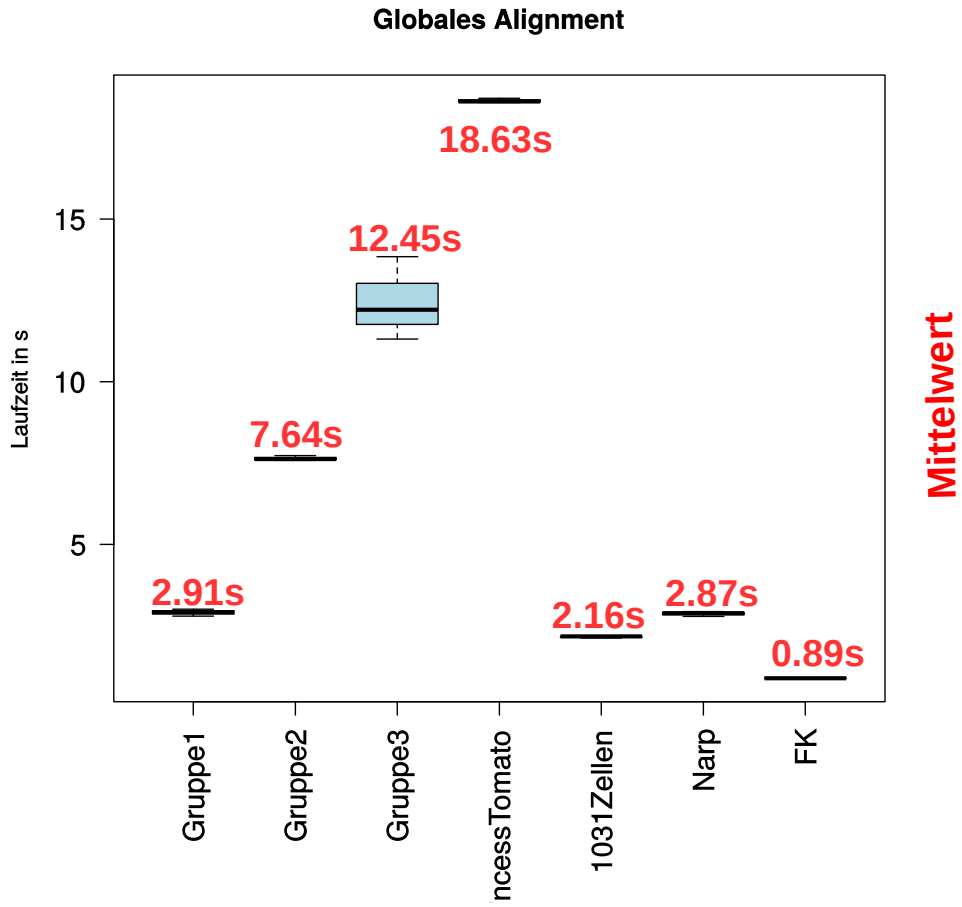
Lösungen/ Wettbewerb Übung 5

Lösungen Übung 5 vorstellen

- Globales Alignment (10P) GruppeAB ?
- Anzahl optimaler Alignments (6P) GruppeXY
- Anzahl möglicher Pfade (4P) GruppeYZ

Wettbewerb Übung 5: Globales Alignment

Ergebnisse aller Gruppen mit richtiger Lösung:



Wettbewerb

	Gruppe0	Gruppe1	Gruppe2	Gruppe3	PrincessTomato	1031Zellen	LL	Narp	FK
Challenge1 (beliebiges Stringmatching)	0	0	0	0	0	3	0	1	2
Challenge2 (Boyer Moore)	0	0	1	1	0	3	0	0	2
Challenge3 (q-grams)	0	0	0	2	0	1	0	0	3
Challenge4 (Suffix-arrays)	0	0	0	0	0	3	0	2	1
Challenge5 (Globales Alignment)	0	1	0	0	0	2	0	1	3
Summe	0	1	1	3	0	12	0	4	11

(Ergebnisse Challenge 5 unter Vorbehalt, bis Korrektur vollständig abgeschlossen)

Übung 6
(Approximatives String-Matching)

Neue Daten

- Auf der Homepage findet ihr „echte“ Wettstreitdaten (String Similarity Search/Join Competition 2013)

- **testdata_reads**
- Beinhaltet 100,000 Sequenzen mit einer Alphabetgröße von 5 (ATGCN)
- Die Länge der Strings sollte zwischen 90 und 110 liegen

- **queries.csv**
- Beinhaltet 499 „Queries“
- Die Länge der Strings sollte zwischen 90 und 110 liegen

- **testdata.csv**
- Beinhaltet einen Goldstandard für die 499 Queries

Übung 6.1

(1) Approximative Suche mit q-gram Index (20P)

- Erweitern Sie Ihren q-gram Index (Aufgabe 3) so, dass eine approximative Suche möglich ist
- Definition: „Given strings s and t, s is k-approximately similar to t, if and only if s can be transformed into t by at most k edit operations. The edit operations are: replacing one symbol in s, deleting one symbol from s, and inserting one symbol into s.“
- Es gilt k={0,4,8,12,16}

- Eine Möglichkeit: Baeza-Yates, Perleberg (BYP)
- Alternativ: Eigene Idee entwickeln

- Inspiration:
  http://www2.informatik.hu-berlin.de/~wandelt/searchjoincompetition2013/Results.html
  Paper: Wandelt, S. and Deng, D. and Gerdjikov, S. and Mishra, S. and Mitankin, P. and Patil, M. and Siragusa, E. and Tiskin, A. and Wang, W. and Wang, J. and Leser, U. (2014) **State-of-the-art in String Similarity Search and Join**. SIGMOD Record, 43 (1).

Format Input (testdata_reads)

- Jede Zeile entspricht einer Sequenz
- **Format: ID, sequence**
  - ID ist ein Identifier für jede Sequenz (Zeile)
  - sequence ist ein String
- Zeilen haben keine Leerzeichen

Format Input (queries.csv)

- Jede Zeile entspricht einer Query
- **Format: ID:query,k**
  - ID ist ein Identifier für jede Sequenz (Zeile)
  - query ist ein String
  - k bestimmt die Anzahl an erlaubten Edit-Operationen
- Zeilen haben keine Leerzeichen

Format Output (1)

- Jede Zeile entspricht einer (Teil)antwort
- **Format: ID:matchId$_0$,matchId$_1$,...,matchId$_n$**
  - ID entspricht der Query-ID
  - MatchId entspricht der ID des matchenden Strings (Template/Index)
- Zeilen haben keine Leerzeichen
- Zeilen enden nicht mit einem Komma
- Zeilenumbrüche zwischen Antworten für Queries

Format Output (2)

- Ergebnisse dürfen sich wiederholen
- Ausgabe von Teilergebnissen ist erlaubt
- Für die Korrektheit der Ergebnisse müssen die beiden Mengen (Goldstandard und Ausgabe) identisch sein.
- Wir stellen den Evaluationscode bereit, welcher zur Entwicklung verwendet werden soll

Beispiel:

```
1:5,11
2:14,10
1:11,12
```

Programmaufruf

- Inputfiles sind unkomprimiert
- Indexierung muss wie folgt aufrufbar sein:

```
java jar Assignment6GrXY.jar index f1 f2
f1 ... Template Datei
f2 ... Indexdatei oder Indexverzeichnis
```

  Ausgabe also in f2

- Suche muss wie folgt ausführbar sein:

```
java jar Assignment6GrXY.jar search F1 F2
F1 ... Indexdatei oder Indexverzeichnis
F2 ... Pattern Datei
```

  Ausgabe auf STDOUT

Wettbewerb

- Das Bauen des Index wird nicht gemessen
  (muss aber innerhalb von zwei Stunden fertig sein!)
- Wir vergeben Punkte für die Schnellste Suche
- Wir messen die Gesamtlaufzeit Ihres Verfahrens für diese Übung (mittels Linux „`time`“, `user+sys`), Parallelisierung lohnt nicht
- Speicherverbrauch oder Größe der Indexdatei ist
  egal solange es auf gruenau2 läuft
- Wettbewerbspunkte:
  - Platz 1: 3 Punkte
  - Platz 2: 2 Punkte
  - Platz 3: 1 Punkt

Abgabe

- Abgabe bis Sonntag den 24.01.2016 um 23:59 Uhr

- Abgabe per Email an: mayeryvo@informatik.hu-berlin.de
  (gerne auch Fragen zur Übung per Email,
  Code Anfragen alte Aufgaben!)

- Jar Datei mit Quellcode (kommentiert!)
  (Dateiname: Assignment6_GrXY.jar, jar ggf. als rar verpacken)

- Kompilierung unter Java 1.7, Jar Datei auf gruenau2 testen,
  Abgaben ohne Quellcode werden ignoriert!

- Bitte die geschätzte Bearbeitungszeit mitteilen
  (z.B. 10h für zwei Teilnehmer)