Grundlagen der Bioinformatik Allgemeines/Übung 1

SS 2016

Yvonne Lichtblau

Allgemeines

Ablauf der Übung

- Insgesamt 6 Übungszettel
- Abgabe in Gruppen von 3 Personen
- Pro Übung ~zwei Wochen Bearbeitungszeit

- 8 Pflichttermine (alle zwei Wochen)
  - Ausgabe neuer Übungszettel
  - Vorstellung der Lösungen letzter Übungszettel

- Termine dazwischen
  - Klärung von Fragen
  - Übungen nach Wunsch (nach Möglichkeit vorher Email)

- Webseite:
  https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/ss16/ue_bioinfo

Termine im Einzelnen

- **22.04.2016**, Ausgabe 1. Übung: Java and FASTA
- **29.04.2016**, Ausgabe 2. Übung: Exact String Matching, Korrektur 1. Übung.
- **13.05.2016**, Ausgabe 3. Übung: Alignments, Korrektur 2. Übung.
- **03.06.2016**, Ausgabe 4. Übung: Hierachical clustering, Korrektur 3. Übung.
- **10.06.2016**, Ausgabe 5. Übung: Introduction to R.
- **24.06.2016**, Korrektur 4. Übung.
- **01.07.2016**, Ausgabe 6. Übung: Microarray Analysis, Korrektur 5. Übung.
- **15.07.2016**, Korrektur 6. Übung.

- Montagstermine folgen
- Ansonsten jeden Freitag/Montag Klärung von Fragen

Übungsschein

- Schein: Voraussetzung für die Prüfung

- Abgabe der Übungszettel in Gruppen von 3 Personen
  - Mindestens 51% der Punkte von jedem Zettel benötigt
  - Gruppen bestehen/scheitern nur als Ganzes

- Vorstellung der Lösungen der letzten Übung durch 2-3 Gruppen
  - Wir losen vorstellende Gruppen aus
  - Ein Student der Gruppe muss Lösung vortragen
    → immer einen Vortrag parat haben
  - Wir behalten uns vor den Student zu bestimmen
  - Ziel: Jeder Student trägt einmal vor

- Klausurtermin: 29.7.2016, 11-14 Uhr

Aufgaben

- Implementationen in JAVA
- I.d.R. sind Eingabe, Ausgabe und Aufrufform vorgegeben
- Textaufgaben
  - Abgabe als PDF
  - Müssen Gruppennnamen beinhalten
- Code-Abgaben
  - Abgabe als Jar (class und source files!)
  - Abgaben ohne Quellcode werden ignoriert
  - Dateiname: UebungX_GrXY.jar
  - Kompilierung unter Java 1.7 (oder niedriger)
  - **Jar Datei auf gruenau2 testen!**
    (gruenau2 ist einer der Rechner hier am Institut, auf den sie sich mit ssh einloggen können: ssh usename@gruenau2.informatik.hu-berlin.de)
- Nichteinhaltung der Regeln: Punkteabzug

Gruppeneinteilung Freitag

Gruppe1:

Gruppe2:

Gruppe3:

Gruppe4:

Gruppe5:

Gruppe6:

Gruppen bitte bis Montag als „GruppeX“ in Goya eintragen!

Gruppeneinteilung Montag

Gruppe1:

Gruppe2:

Gruppe3:

Gruppe4:

Gruppe5:

Gruppe6:

Gruppen bitte bis Montag als „GruppeX“ in Goya eintragen!

Übung 1
Java und FASTA

FASTA Format

- Für die Aufgabe müssen Sie eine Datei einlesen, die Protein Sequenzen enthalten

- Die Datei ist im Fasta Format
  - Zitat: „A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence by a greater-than symbol („>“) in the first column. … The sequence ends if another line starting with a „>“ appears; this indicatesthe start of another sequence“
  - Beispiel:

```
> gi|5524211|gb|AAD44166.1| cytochrome b
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMS
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPL
LLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVP
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQP
>gi|5454351|gb| cytochrome x
LLLITMATAFMGYVLPWGQMSLCLYTHIGRNIYYGSYLYSETWNTGIM
LLLITMATAFMGYVLPWGQMS
```

Aufgabe 1

(1) Einlesen von Sequenzen (10 Punkte)

- Auf der Homepage ist eine Datei bereitgestellt, die Protein Sequenzen im FASTA Format beinhaltet:
  - chr21.fa
- Benutzen Sie Java um die Sequenzen zu laden
- Schreiben Sie für jede Sequenz in der Datei den Header und die Länge der Sequenz raus (tsv Format, tab-separated values):

```
gi|5524211|gb|AAD44166.1| cytochrome b<tab>188
gi|5454351|gb| cytochrome x<tab>70
```

Abgabe

- Abgabe bis Mittwoch den 27.04.2016 um 23:59 Uhr
- Abgabe per Email an: yvonne.lichtblau@informatik.hu-berlin.de (gerne auch Fragen zur Übung per Email)
- Abgabe:
  - Lösung im tsv-Format
  - Code als Jar Datei wie beschrieben (siehe Folie 6)