Assignment 2: Sequence Alignment

SS 2018

Grundlagen der Bioinformatik

Organisation

- Alle Gruppen gebildet?
- Alle Gruppen da?
- Dank an aktive Studentinnen und Studenten
  - Herr Vogt
  - Frau Bitner
  - Herr Choi

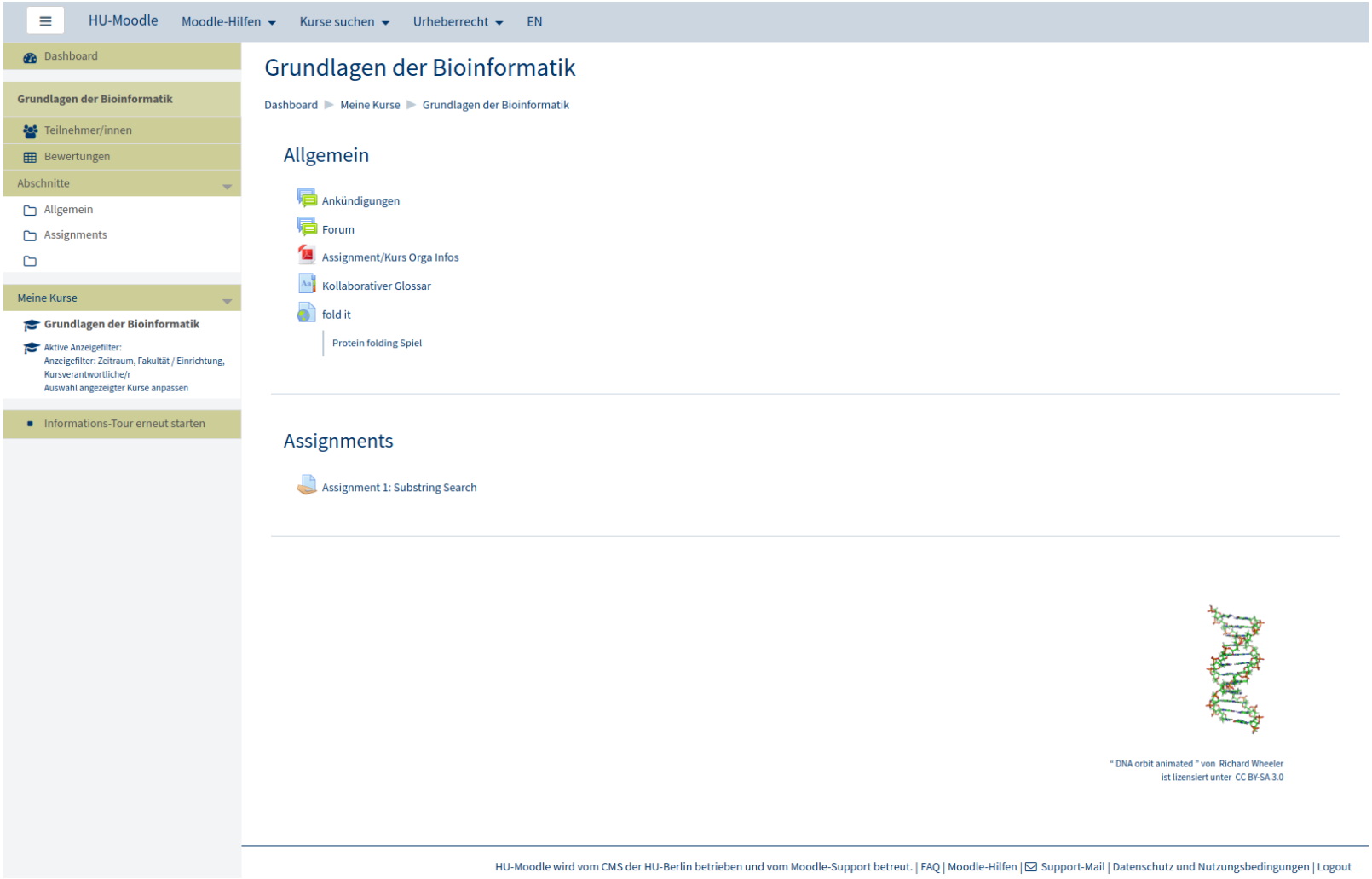
Einsicht Assignment Bewertungen

Nach der morgiger Übung hier Online einsehbar

https://docs.google.com/spreadsheets/d/1QFX_HjWtCUZIv2VGxCFnnEdli56rrAcYvs_C7zBrKDY/edit?usp=sharing

Wenn öffentliche Gruppen-Ergebnisse <u>nicht</u> erwünscht sind
→ Email

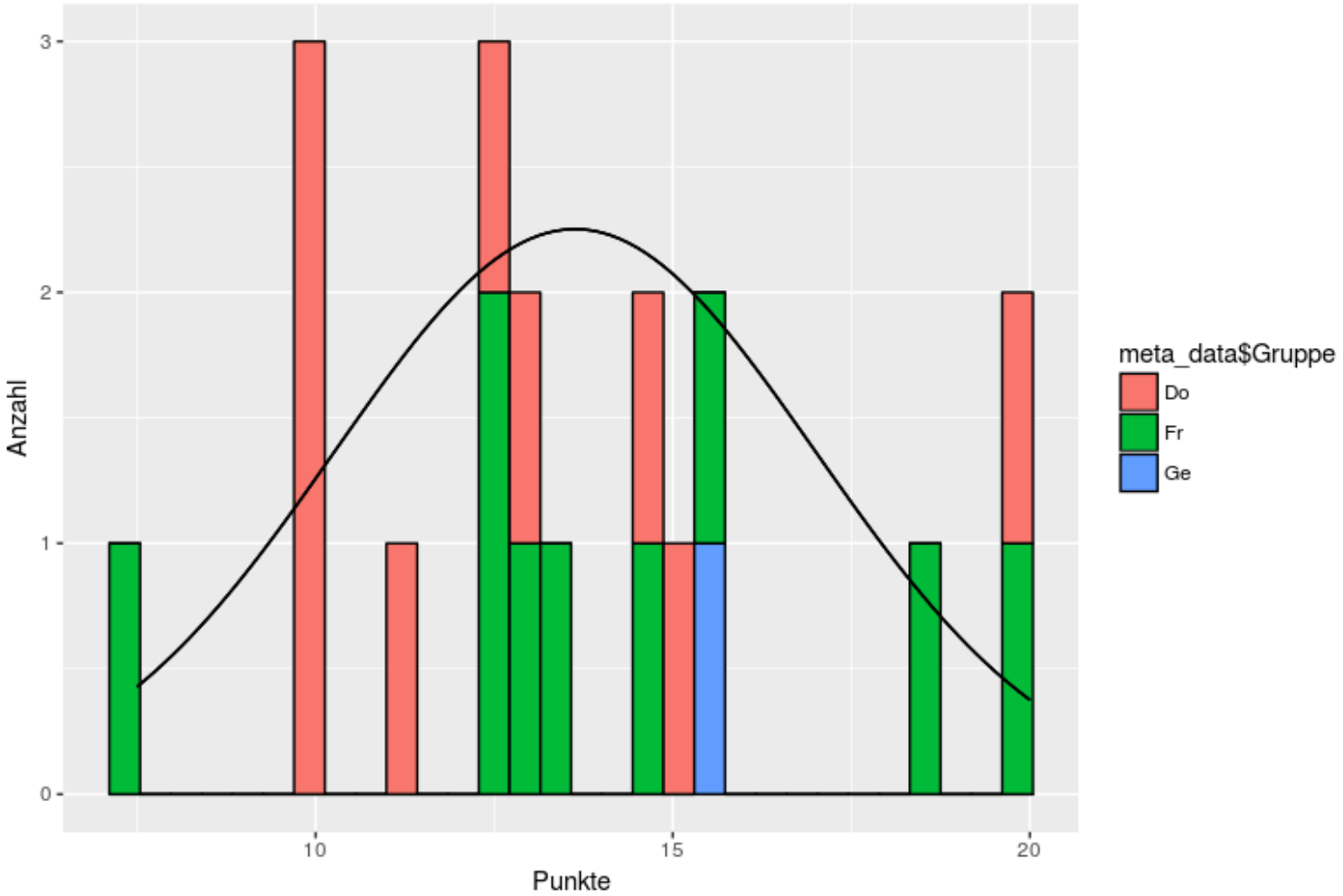
GdB Moodle


HU-Moodle
Moodle-Hilfen
Kurse suchen
Urheberrecht
EN
Dashboard
Grundlagen der Bioinformatik
Teilnehmer/innen
Bewertungen
Abschnitte
Allgemein
Assignments
Meine Kurse
Grundlagen der Bioinformatik
Aktive Anzeigefilter:
Anzeigefilter: Zeitraum, Fakultät / Einrichtung, Kursverantwortliche/r
Auswahl angezeigter Kurse anpassen
Informations-Tour erneut starten
Grundlagen der Bioinformatik
Dashboard ▶ Meine Kurse ▶ Grundlagen der Bioinformatik
Allgemein
Ankündigungen
Forum
Assignment/Kurs Orga Infos
Kollaborativer Glossar
fold it
Protein folding Spiel
Assignments
Assignment 1: Substring Search
" DNA orbit animated " von Richard Wheeler
ist lizensiert unter CC BY-SA 3.0
HU-Moodle wird vom CMS der HU-Berlin betrieben und vom Moodle-Support betreut. | FAQ | Moodle-Hilfen | Support-Mail | Datenschutz und Nutzungsbedingungen | Logout

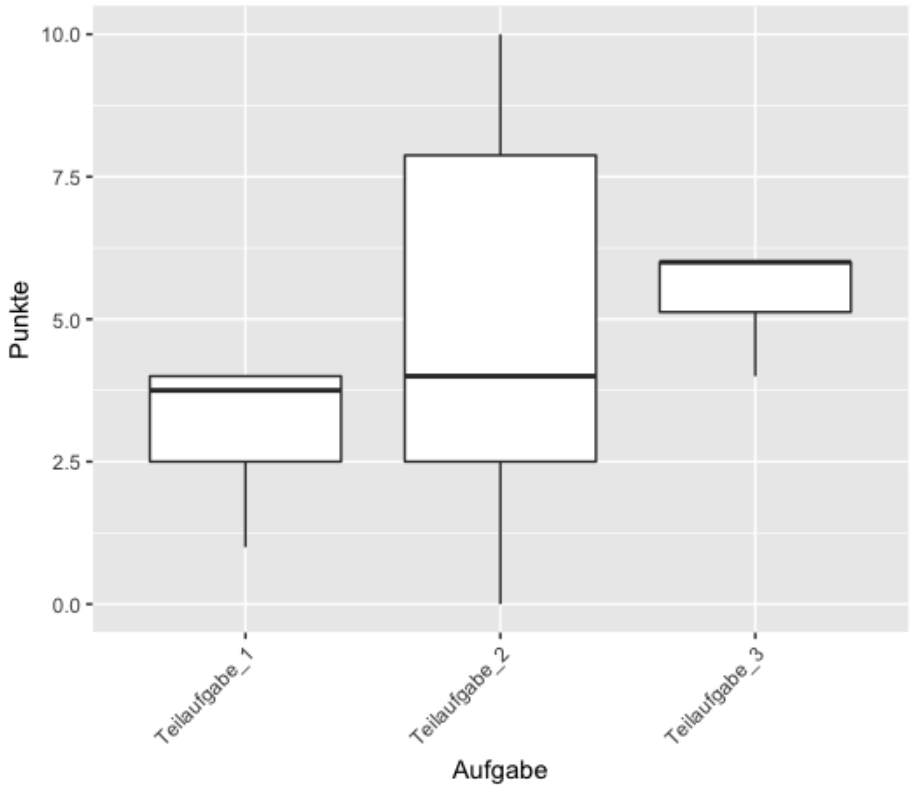
Aufgaben

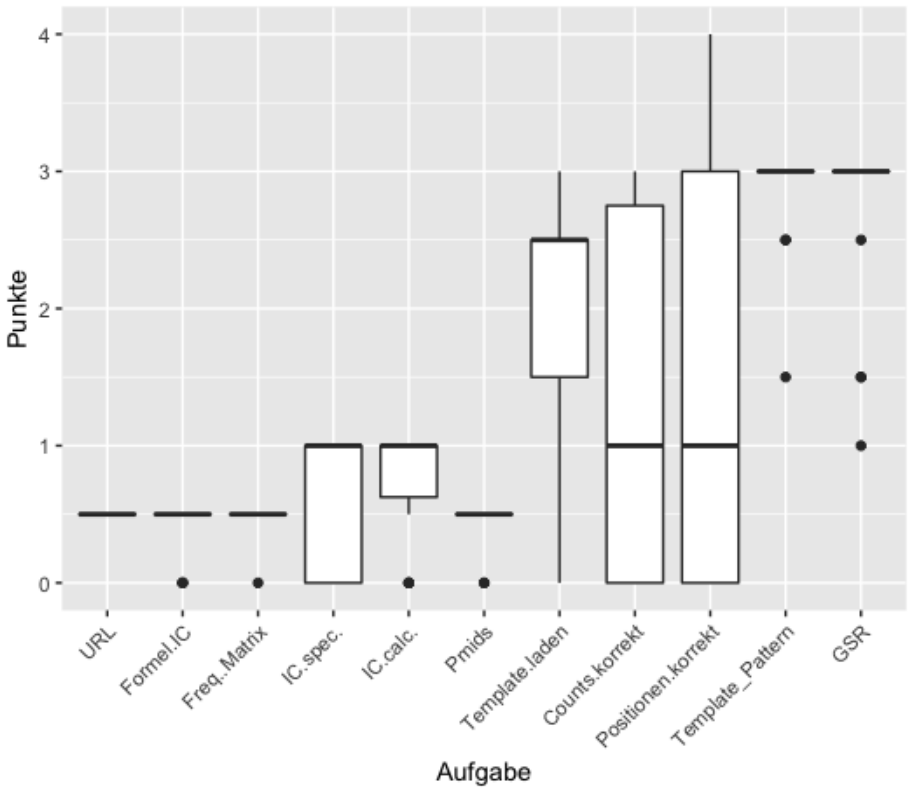
1. Analyze Transcription Factor GATA2 (4 p)
2. Substring Search (10 p)
3. Properties of Boyer Moore Algorithm (6 p)

Histogram + Normerteilungsfit von aller Gruppenpunkte

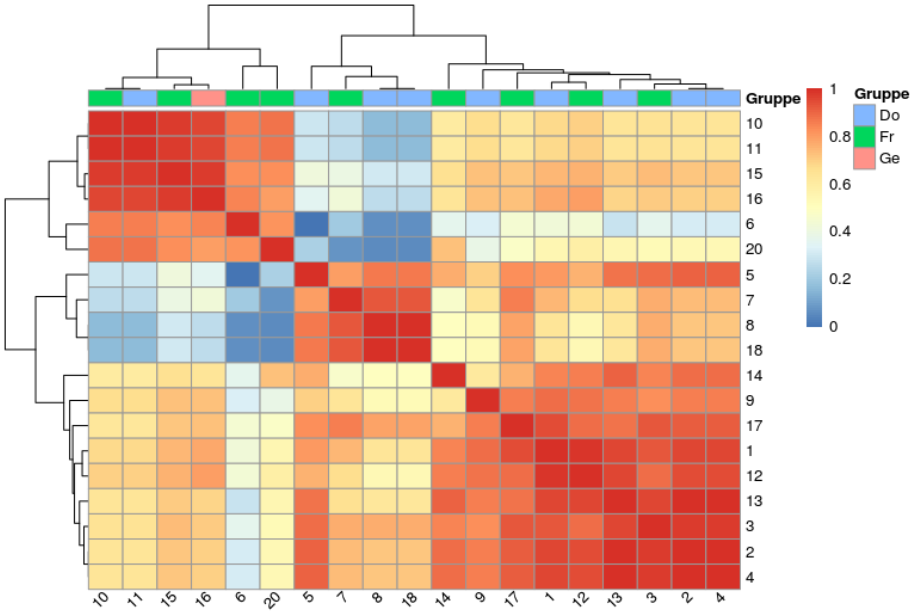


Boxplots Punkte pro Teilaufgabe bzw. pro Unterteilaufgabe

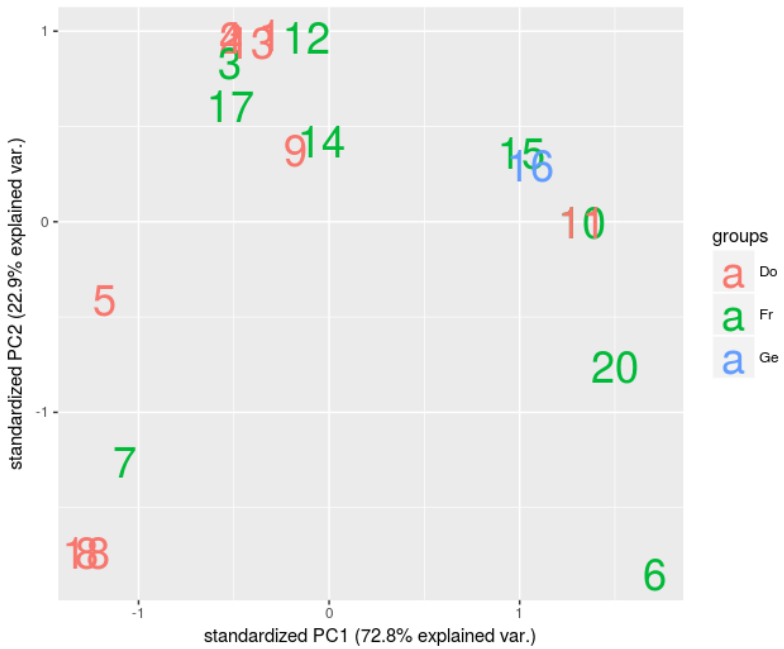




Ähnlichkeiten der Gruppenergebnisse, gefärbt nach Tag



Correlation Heatmap



Pca of correlation matrix

Assignment 2
Sequence Alignment

Overview – Assignment 2 (20P)

(1) Local Alignment (10P)

(2) Global Alignment (5P)

(3) Aligning real sequences (5P)

(1) Local Alignment (10P)

- Write a program to compute the local similarity of two DNA sequences using Smith Waterman
  - Sequences must be read from a FASTA file (pair.fasta) (1P)
  - Use replacement costs provided in matrix file (matrix.txt) (2P)
    - Deletion/Insertion cost is 8
  - Print length of best local alignment, score AND number of
    - Matches
    - Replacements
    - Deletions
    - Insertions (3P)
  - Print alignment (4P)

```
AAATT_GCC
|. || |.|
AC_TTTGGC
```

(1) Global/Local Alignment (10P)

Global alignment

		A	T	G	T	C	G
	0	−1	−2	−3	−4	−5	−6
A	−1	1	0	−1	−2	−3	−4
T	−2	0	2	1	0	−1	−2
G	−3	−1	1	3	2	1	0

```
ATGTCG
ATG____

ATGTCG
AT____G

ATGTCG
A___T_G
```

Local alignment

		A	T	G	T	C	G
	0	0	0	0	0	0	0
A	0	1	0	0	0	0	0
T	0	0	2	1	1	0	0
G	0	0	1	3	2	1	1

```
ATG
ATG
```

Daten

Notwendige Dateien können von hier bezogen werden

https://box.hu-berlin.de/d/db067889ea924927aacd/

- Programmaufruf:
  `java -jar GdBioinf[ÜbungNr]_[Gruppe].jar pairs.fasta matrix.txt`

  - **Tipp: Score für Task 1 ist zwischen 150 and 170!**

(1) Local Alignment (10P)

pair.fasta:

```
>seq1
CCCAGCAGCAGAAGTTATCACTGGCTATCAACGATTGAACTCCCAATGTGGCGAGCAACGGA
CGGCACAGCAGGCAGCCTTACTCCATGTTGTTCGACAATACTCAGTTCTACAGTCCAG
>seq2
CTGAGCACCGCTTTTGCACTACAAGGATTCGAACCCCATTGTGCGAACAACGGACGCACAGC
ATTACACCTGTTTGCCGATATTCACCCTGATGTGGG
```

matrix.txt:

```
#
# DNA scoring matrix
#
# Lowest score = -4, Highest
score = 5
#
    A   T   G   C
A   5  -3  -4  -4
T  -3   5  -4  -4
G  -4  -4   5  -2
C  -4  -4  -2   5
```

deletion/insertion cost is 8
→ score = -8

(2) Global Alignment (5P)

Derive a formula which calculates how many optimal alignments exist between strings *n* and *m*.

$|n| > |m|$

<u>One-element alphabet</u>
(e.g. only ‚A‘)

Explain how you derived this formula.

(3) Aligning real Sequences (5P)

- *KRAS* is a *RAS* family member and an important oncogene. Mutation status is used to estimate drug response for colorectal cancer
- Download the DNA sequences for human (NM_004985.3) and mouse (NM_021284.6): www.ncbi.nlm.nih.gov/nuccore
- Calculate local alignment score and alignment using your program (1P)
- Calculate local alignment score using EMBOSS (2P)
- Are the results the same? Discuss if not. Explain the required steps to get the same results (2P)

(3) Aligning real Sequences (5P)

EMBOSS

- **E**uropean **M**olecular **B**iology **O**pen **S**oftware **S**uite

- Framework for many tasks
  - Sequence retrieval
  - Alignment
  - Folding
  - Motif finding
  - ...

- Can be used online or locally
  - http://emboss.sourceforge.net/
    http://emboss.bioinformatics.nl/

(3) Aligning real Sequences (5P)

EMBOSS http://emboss.bioinformatics.nl/

[sort alphabetically]

ALIGNMENT
extractalign

ALIGNMENT CONSENSUS
cons
consambig
megamerger
merger

ALIGNMENT DIFFERENCES
diffseq

ALIGNMENT DOT PLOTS
dotmatcher
dotpath
dottup
polydot

ALIGNMENT GLOBAL
est2genome
needle
needleall
stretcher

ALIGNMENT LOCAL
matcher
seqmatchall
water
wordmatch

EMBOSS explorer

Welcome to EMBOSS explorer, a graphical user interface to the EMBOSS suite of bioinformatics tools.

To continue, select an application from the menu to the left. Move the mouse pointer over the name of an application in the menu to display a short description. To search for a particular application, use wossname.

For more information about EMBOSS explorer, including how to download and install it locally, visit the EMBOSS explorer website.

Development of EMBOSS explorer has been supported by the National Research Council of Canada and Genome Prairie.

Abgabe

- Abgabetermin 30.05.2018 um 23:59 Uhr

- Fragen: raik.otto@hu-berlin.de
- Abgabe als .zip hier hochladen:
  https://box.hu-berlin.de/u/d/5126a9e3a7/
  - Dateiname: GdBioinf_2_Gruppe_[Gruppennummer].zip

- .jar/.py/.R auf gruenau2 testen!

Was abzugeben ist

- Abgabetermin 30.05.2018 um 23:59 Uhr

- PDF mit
  - Teilaufgabe 1: Output eures Programms
  - Teilaufgabe 2: Antwort
  - Teilaufgabe 3: Output Eures Programms, Emboss Score, Antwort zu Task 3
- Code als Jar Datei (
  - Dateiname: GdBioinf_[Assignment_nr]_Gruppe_[GruppenNR].jar)
- Sourcecode (einzeln in der Zip-Datei oder mit in der Jar-Datei)