Grundlagen der Bioinformatik
Assignment 2: Sequence Alignment

Yvonne Lichtblau
SS 2017

Vorstellung Lösungen Übung 1

Overview – Assignment 1 (20P)

(1) Analyse transcription factor *GATA2* (4P)
Eine Gruppe

(2) Substring search (10P)
Zwei Gruppen

(3) Properties of Boyer Moore Algorithm (6P)
Eine Gruppe

Assignment 2
Sequence Alignment

Overview – Assignment 3 (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, number of matches, replacements and deletions+insertions (3P)
  - Print alignment (4P)

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

  - Programmaufruf:

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

(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
```

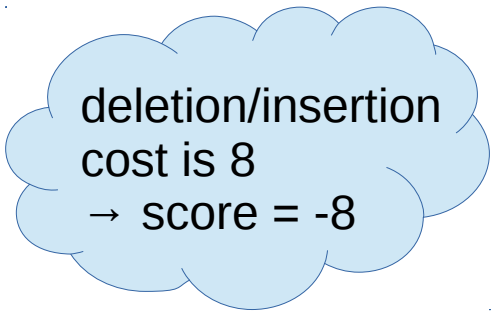
(1) Local Alignment (10P)

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
```



(2) Global Alignment (5P)

Derive a formula which calculates how many optimal alignments exist between a string of length *n* and a string of length *m*, if both strings are defined over the same one-element alphabet.

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

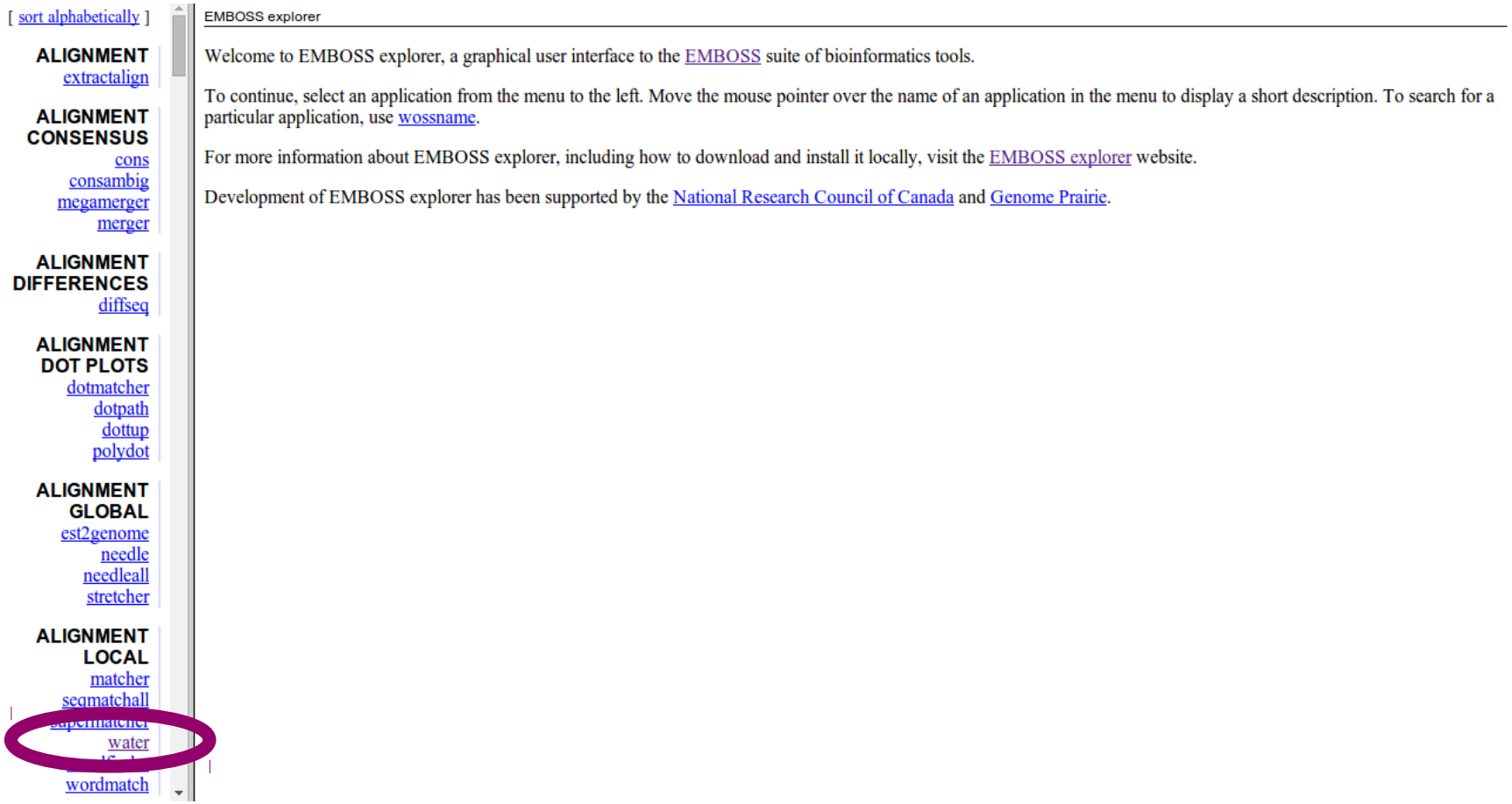
- **E**uropean **M**olecular **B**iology **O**pen **S**oftware **S**uite

- 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/



Abgabe

- Abgabe bis Dienstag den 30.05.2017 um 23:59 Uhr

- Fragen zur Übung gerne per Email: yvonne.lichtblau@informatik.hu-berlin.de
- Abgabe als .zip hochladen:https://box.hu-berlin.de/u/d/5126a9e3a7/
  - Dateiname: GdBioinf[ÜbungNr]_[Gruppe].zip (z.B. GdBioinf2_Gruppe2.zip)
  - PDF mit
    - Task 1: Output eures Programms
    - Task 2: Antwort
    - Task 3: Output Eures Programms, Emboss Score, Antwort zu Task 3
  - Code als Jar Datei (Dateiname: GdBioinf[ÜbungNr]_[Gruppe].jar)
  - Sourcecode (einzeln in der Zip-Datei oder mit in der Jar-Datei)
- .jar auf gruenau2 testen!

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