Grundlagen der Bioinformatik

Exercises - Introduction

Ulf Leser

Assignment 1 - Overview

- Read a (large) DNA file in FASTA format
- Implement and run Boyer-Moore for exact pattern matching
- Study (the basics of) sequence logos for TFBS

Background: FASTA

- Extremely common, human-readable text file format for storing / exchanging sequences (DNA, RNA, protein)
  - Human-readability is often much more important than compactness
- Definition
  - *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 data by a greater-than (">") symbol in the first column. ... The sequence ends if another line starting with a ">" appears; this indicates the start of another sequence.*"
- Example

>gi|5524211|gb|AAD44166.1| cytochrome b
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSEWIWGGFSVDKATLNRFFAFHFILPFT
MVALAGVHLTFLHETGSNNPLLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPGLMPFLH
TSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQP
>gi|5454351|gb| cytochrome x
LLLITMATAFMGYVLPWGQMSLCLYTHIGRNIYYGSYLYSETWNTGIM LLLITMATAFMGYVLPWGQMS
>gi …

Task 1

- In moodle, you find two files
  - sequence.fasta (~50MB)
  - patterns.fasta (~100B)

- Write a program that can read FASTA files
  - Test (at least) with both files
- Output: Length of all sequences in file
- Submission: program called "**fastaread**" with one parameter: Name of FASTA file
  - Need not check whether the file actually is in FASTA format

Expected input / output

- Input file

```
>gi|5524211|gb|AAD44166.1| cytochrome b
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSEWIWGGFSVDKATLNRFFAFHFILPF
TMVALAGVHLTFLHETGSNNPLLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPGLMP
FLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQP
```

- Output (nothing else)
  - 187

- Input file

```
> pat
LLLI
> pat
LAGVGGF
```

- Output (nothing else)
  - 4
  - 7

Task 2: Implement Boyer Moore

- Write a program that reads two FASTA files and searches all sequences (patterns) of the second file in the first
  - First file contains only one sequence
  - Of course: Re-use code from previous task
  - Search must use your own implementation of Boyer-Moore
- Submission: program called "**boyermoore**" with two parameters (FASTA files)
- Output: For every pattern in second file
  - Number of times the Bad Character Rule was applied
    - Implement BM as presented in the lecture
    - If GSR and BCR find the same shift: Also count as BCR
  - Number of occurrences of pattern
  - Positions of the first 10 occurrences (or less if there are less)

Expected input / output

- Input file 1

```
>gi|5524211|gb|AAD44166.1| cytochrome b
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSEWIWGGFSVDKATLNRFFAFHFILPF
TMVALAGVHLTFLHETGSNNPLLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPGLMP
FLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQP
```

- Input file 2

```
>…
LLL
>…
LLI
>gi|5454351|gb| cytochrome x
LLLQQQ
```

- Output (nothing else, this format)
  - ? / 5 / 29 / 96 / 97 / 98 / 99
  - ? / 1 / 31
  - ? / 0

Competition

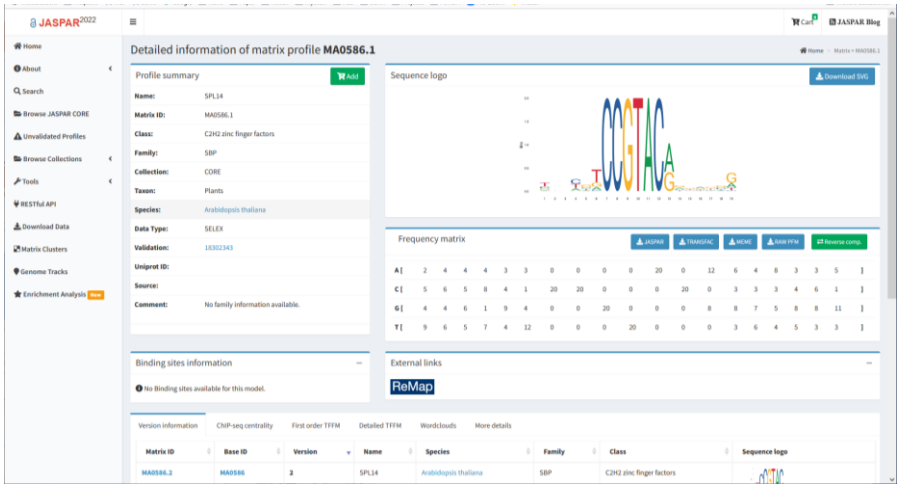
- Provide a fast version of Boyer Moore!
  - Must use Boyer-Moore or variant – we will check
  - Can be different from solution to Task 2 (different counts for BCR)
  - Submission (voluntarily): program called "**boyermoore_competition**" with two parameters (FASTA files)
- We will measure wall clock time (unix time) for new combinations of sequence / patterns
  - Alphabet is {A,C,G,T,N}
- Example from previous editions
  - Fastest solutions: ~1.1sec
  - Slowest solutions: >300sec

Things you may consider for speeding-up search

- Faster preprocessing
- GSR with / without previous character
- Faster IO
- If there are only X characters left to compare – is computing shifts of GSR and BCR worth it?
- If BCR suggests a "long" shift – worth to still compute GSR-shift?
- Variants of GSR / BCR
- …

Task 3: TFBS and Sequence Logos

- Search the JASPAR database
- Search the TF GATA2 (1st version, MA0036.1)
- Compute the information content of each position in the position specific weight matrix (PSWM)
- Identify the correct formula for the information content yourselves



Submission (as PDF)

- URL to the JASPAR entry
- Formula for information content used in sequence logos
- Frequency matrix and IC for every position in the PSWM
  - Add a table to the PDF
- List three cancer types that GATA2 is associated with
  - Use PubMed
  - Give name of cancer, title of publication, and PMID
  - To need for extensive checks of the strength of the association found in the paper

General requirements

- Remember to name all programs as requested
- All programs must run without further installations on GRUENAU2
  - *ssh username@gruenau2.informatik.hu-berlin.de*
- For all programs, source code must be submitted as well
  - Document your code
  - For Java/C etc.: Submit the source code and the compiled binary
- All responses must be submitted as PDF, where the task /assignment of every answer is clearly recognizable
- Zip everything into one file and upload via Moodle
  - **AssignmentX_groupY.zip**
- Deadline for submissions of assignment 1: 12.5.2022, 11:00

Questions?