Exercises to Introduction to Bioinformatics
Assignment 1: Substring Search

Philippe Thomas
Tasks

1. Analyze Transcription Factor GATA3 (5 p)

2. Substring Search (9 p)

3. Properties of Boyer Moore Algorithm (6 p)
1. Transcription Factor GATA3 (5 p)

- **GATA3** is a transcription factor with established or assumed roles in a variety of different human cancers
- Search GATA3 in the JASPAR database

- **JASPAR** contains a set of transcription factor DNA-binding preferences, modeled as matrices
- Profiles derived from published collections of TF-binding sites
- Profile can be used to scan query sequences for presence of potential binding sites
1. A Real Transcription Factor (5 p)

- Search GATA3 in the JASPAR database
- Compute the information content of each position in the PWM
  - Find the exact formula on the web

- What I want
  - Link to the JASPAR information on GATA3 (1p)
  - Formula for information content used in sequence logos (1p)
  - Frequency matrix and IC of every position of the PWM (1p)
  - List of cancer types to which GATA3 is associated and supporting papers from PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) (2p)
2. Substring Search (9 p)

- Load a template string (70MB) into main memory
- Load a set of patterns
- Search all exact occurrences of all patterns in the template
2.1 Load a Sequence (3 p)

- You need to load sequences in **FASTA format**
  - „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.“
  - **Handle characters as they are** (‘ATGC’ but also ‘WBN..’)
  - Example
    ```
    >gi|5524211|gb|AAD44166.1| cytochrome b
    LCLYTHIGRNIYYGYSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMS
    EWIWGGFSVDKATLNRFFAFHFIPLFMTMVALAGVHLTFLHETGSNNPL
    GLMPFLHTSKHRSMLRPLSQAFLFWTLTLTLTLTLTWIGSQP
    >gi|5454351|gb| cytochrome x
    LLLITMATAFMGYVLPWGQMSLCLYTHIGRNIYYGYSYLYSETWNTGIM
    LLLITMATAFMGYVLPWGQMS
    >gi ...
    ```
2.2 Load a Set of Patterns (0p)

- You will get another file which contains a set of sequences in FASTA format. These should be used as patterns.
- See pattern_aufgabe.fasta on website
2.3 Substring Search (6p)

- Implement an **algorithm of your choice** to search all occurrences of all patterns in the template
- **Note:** STRING.indexof() or the like **is not accepted**
- Some *approximate* frequencies (for checking)
  - tccgga: 2506 (this is the only exact frequency)
  - gctacc: 6200
  - taataa: 22700

- **What we want**
  - A Java Archive including class files and source code
    - java –jar Assignment1-grXY.jar patternFile.fasta templateFile.fasta
  - List of patterns with number of occurrences
    - >tccgaa 2506; >gctacc XY; ...
  - Runtime of the algorithm and short description of machine where it run and a sentence on the implementation (Boyer Moore, naive, ..)
3. Properties of the Boyer Moore Algorithm (6 p)

• Give a template and a pattern such that the BM algorithm, as presented in the lecture, needs to do in the order of $|T|*|P|$ comparisons of characters (3 points)
  – And explain why

• Many implementations of the BM algorithm actually drop the good suffix rule, especially for larger alphabets. Give an argument why this can be useful (3 points)
Submission

• Submit all requested data as plain text by Wednesday 2.05.2012, 23.59
• Approximate time needed to complete assignment
• Send by mail to: thomas[at]informatik.hu-berlin.de