Exercises to
Introduction to Bioinformatics
Assignment 1: Substring Search

Ulf Leser
1. A Real Transcription Factor (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
- Compute the **information content** of each position in the PSWM
  - Find the exact formula on the web
- **What we want**
  - Link to the JASPAR information on GATA3 (1p)
  - Formula for information content used in sequence logos (1p)
  - IC of every position of the PSWM (1p)
  - List of cancer types to which GATA3 is associated and supporting papers from Medline (2p)
2. Substring Search (9 p)

- Load a (long) template 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.
  
  - Example
    
    >gi|5524211|gb|AAD44166.1| cytochrome b
    LCLYTHIGRNIYYGSYLYSETWNTGIMLLLLITMATAFMGYVLPWQMS
    EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPL
    LLLLALLSPDMLGDPDHMPADPLNTPLHIDKPEWYFLFAAYAILRSVP
    GLMPFLHTSKHRSMLRPLSQAFLWTLTMDLLLTNTWIGSQP
    >gi|5454351|gb| cytochrome x
    LLLITMATAFMGYVLPWQMSLCLYTHIGRNIYYGSYLYSETWNTGIM
    LLLITMATAFMGYVLPWQMS
    >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.
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 exact frequency)
  - gctacc: 6200
  - taataa: 22700
  - cccccccccccc: 50
- **What we want**
  - The program in source code
  - List of patterns with number of occurrences
  - Runtime of the algorithm and short description of machine where it run
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 Sunday, 14.11.2010, 23.59
• Send by mail to Prof. Leser