Sequence Bioinformatics: Motif Searching and Visualization

The purpose of this project is to investigate regulatory motif occurrences in yeast promoters. Transcription factors are proteins that regulate gene expression of target genes by binding to specific sequences in regions upstream of the transcription start site. Binding of the transcription factors to the regulatory sequences can either induce or repress transcription of the target genes.

The sequences (also called motifs) that transcription factors bind to can be described in different ways. The difficulty in identifying and describing these sequences is that the sequences often are degenerate, i.e. that certain positions in the motif can be variable. Most often, the motifs are described with position weight matrices that indicate which nucleotides are possible at the different positions of the motif. A motif can also be illustrated with its consensus sequence.

Your assignment is two-fold:

> (i) Identify a set of possible binding sites in a selection of yeast sequences (upstream regions) by using regular expressions in Python. The motifs shall be described by regular expressions that allow for variability within each position if it is needed (but we do not care about the different weights for nucleotides at the positions).
> (ii) Visualize these binding sites by a simple program in Java (a small program will be provided that you can modify).

To pass the project, at least part (i) should be solved, and the output given in simple text format.

Hints:

> (i) Sequences in python can be contained within a list or a dictionary. Some tips on getting started with Python can be found here: http://python.genedrift.org/2007/10/10/alternative-methods-to-split-a-fasta-file/.
> (ii) Regular expressions are very powerful to identify patterns in strings. In Python this is achieved by using the re module. Some hints on how to use regular expressions with Python is given here: http://docs.python.org/dev/howto/regex.html.
> (iii) Graphics is relatively easy with Java. However, if you have no previous Java experience, it might still be tricky to get started. Java tutorials can be found at http://download.oracle.com/javase/tutorial/. If you have time, learning some Java for part (ii) can be fun.

Files:

sequences.fasta : contains a set of yeast sequences with 500 bp upstream and 50 bp downstream of the transcription start site for selected genes.

motifs.txt : Contains four motifs, three of which are described with weight matrices. Your job is to translate these matrices into regular expressions and search for them within the sequences.

output.pdf : an example of a simple graphical output. There is no need to have line numbers in your output.

Good luck!