Lab 4 – Algorithms

To get the lab approved, send your answers to: david.sundell@plantphys.umu.se

Task 1 – Time complexity

What is the time complexity of this algorithm (from Lab 1):

```
RemoveDuplicates(list, n)
1       newlist ← ()
2       for i ← 1 to n
3           m ←length of newlist
4           foundDuplicate ← false
5           for j ← 1 to m
6               if list_i = newlist_j
7                       foundDuplicate = true
8                       break
9           if foundDuplicate = false
10              add list_i to newlist
```

Task 2 – Approximation algorithms

Assume that you have written a minimization algorithm A that outputs 15 for some input $\boldsymbol{\pi}_1$ (i.e. A($\boldsymbol{\pi}_1$) = 15) and 5 for another input $\boldsymbol{\pi}_2$ (i.e. A($\boldsymbol{\pi}_2$) = 5). Furthermore, assume that you have run an exhaustive search for these two inputs and thus know the optimal output: OPT ($\boldsymbol{\pi}_1$) = 10 and OPT ($\boldsymbol{\pi}_2$) = 2. What can you say about the performance guarantee of algorithm A? Explain.

Task 3 – Motif search

The file promoters.txt contains some Yeast promoters in fasta format. According to the SCPD database (http://rulai.cshl.edu/SCPD/) these genes are regulated by the cell cycle related transcription factor MCM1 through a 10mer with consensus sequence CCNNNWWRGG:

W	= A or T	S	= C or G
R	= A or G	Y	= C or T
K	= G or T	M	= A or C
B	= C, G, or T	D	= A, G, or T
H	= A, C, or T	V	= A, C, or G
N	= A, C, G, or T		

There are a myriad of motif finding tools available: MEME, CisModule, AlignAce, PhyloGibbs, Weeder, Amadeus, FIRE, SCOPE, etc. We will concentrate on MEME here since it is commonly used, has good documentation and nice visual output.

MEME can be run from here: http://meme.sdsc.edu/meme/cgi-bin/meme.cgi. Look into the documentation, especially the parameter available for tweaking, and answer the following questions:

What type of algorithm does MEME use (i.e. which algorithm design technique from the slides)?

Which motif distribution (i.e. one, zero or one, or any number per sequence) would you use for the yeast promoters above?

Which motif width would you use?

Run MEME with your preferred settings. Leave the other parameters unchanged. Even on this small example, the algorithm takes several minutes. To not waste time, there is a link to the MEME-output file on the course web page. Explore the results (i.e. "MEME output as HTML") and answer the following questions:

Do any of the found motifs look like the known consensus motif CCNNNWWRGG?

Does the position of the found motifs in the promoters seem to matter?

STAMP (http://www.benoslab.pitt.edu/stamp/) is a tool for comparing a newly found motif to databases of known motifs. Under "Data Formats" in the MEME results, select "Raw Format" and past the result into the STAMP "Input Motifs"-window. Add, for example, ">Possible MCM1-motif" as the first line (if not, STAMP will not accept the input). Press "Submit". What is the closest known motif?

Meme finds motifs in the input sequences that are unlikely to occur there by chance (as indicated by the p-values). However, MEME do not know what words exist in the genome at large (the entire Yeast genome in our case). In fact, MEME only knows the background model, which is the distribution of A, G, T and C in the nr database (or in a user-specified database). Analogously, if we knew the distribution of letters in a book, and were given the task to find unlikely words on ten specific pages, we might choose a word like "jalapeño" if it occurred on all these pages. But, what if this was a cookbook with recipes using jalapeños? It is likely that all pages in this book would contain the word "jalapeño" without affecting the background distribution much. With this analogy in mind, explain how the option "Perform discriminative motif discovery – Enter the name of a file containing 'negative sequences'" in MEME can solve this problem.

How do we know that the E-value of the motif (and the p-values of the individual occurrences in the promoters) is significant? At the course web page there is a link to

the MEME-output after checking the box “Shuffle sequence letters”. What is your conclusion after looking at these results?