Exercises „Introduction to Bioinformatics“

Part 1:

1.1
Is there a problem with the following algorithm when the variable i reaches N + 1? (Hint: what is a[i] in this case?)

```
i=1;
while (i<=N && a[i]==b[i]) { i=i+1;}
if (i>N) {equal = true;} else {equal = false}
```

1.2
Assume that the hypothetical amino acids A R N and D occur at similar frequencies in genes. Given the following two examples of good matches, what substitution matrix would result?

A R R N D R N A D D N R A N D N D N N N R D A D A R N R D A A D A A N
A R D N D R A D A D N R A D D A D A N N R D A A D R N N D A A N A A N

D D N D N D N A N A D A D A N N D A N D A N D N A N D D N R R R
D A D D N N N A N A D A N A N N N A N D A N N N A N R A A R D R

1.3
Given the following substitution matrix, score the following alignment

	A	R	N	D
A	5			
R	-2	7		
N	-1	-1	7	
D	-2	-2	2	8

```
DRDANND
ARDNRND
```

1.4
Write a program that finds a substring in a string (on a character by character basis, not using built-in functions). Do this either in Java (if you can) or in Excel. Hint: with Excel, you have to write iterations/loops spatially, i.e. use one cell per iteration and copy the body of the loop into each cell.

1.5
Write down the algorithm for overlay matches.

1.6
BLAST: Given the substitution matrix above and the search pattern R R D N A R D, list all four character words with an alignment better than 3.

1.7
Assume that a splice site is described by the pattern { C } * (A | G | T) G G T (A | G) . Which of the following are splice sites?
A. AGGTA
B. CGGGTG
C. CCTGGTC
D. CCCCCCGGGTG

1.8
Write down the state table for the deterministic finite automaton for the pattern (A | C) D.

1.9
Given the following 5 measurements and their relative distances given by the following matrix, build a tree by UPGMA clustering.

	A	B	C	D	E
A	0	9	10	17	19
B		0	10	17	17
C			0	8	10
D				0	4
E					0

Part 3:

3.1
Do a segmentation of the following bitmap with a threshold of 50 (numbers indicate pixel intensities):

00 12 13 15 13 08
02 13 22 23 17 04
05 13 51 58 27 12
04 49 70 80 55 36
00 43 60 80 40 22
00 12 20 30 20 24

3.2
If the pixels represent fluorescence intensities, calculate total intensity and average intensity of the segmented object.

3.3
Retrieve the appropriate data from the diauxic shift experiment from deRisi et al. (http://cmgm.stanford.edu/pbrown/explore/).

3.4
Write a small clustering program (you can probably also simulate a simple clustering algorithm with Excel). Print the expression matrix ordered according to the clustering (either numerically or graphically, if you can). Vary clustering parameters and observe differences in the result. Compare your result with those from the paper.

3.6
Look up the enzymes from the metabolic chart in SwissProt (http://www.expasy.org/) and describe what additional information you get through those records.

3.7
Look up the corresponding genes and see whether there are parallels/differences of these pathways in other organisms.

3.8
Obtain reactions entries from the KEGG database (http://www.genome.ad.jp/kegg/kegg2.html), summarize the contents and try to construct a pathway diagram through the KEGG-provided or other tools on the web.