CHALMERS UNIVERSITY OF TECHNOLOGY

Examination in Bioinformatics, MVE360

Saturday 3 March 2012, 14:00-18:00

Examiner:	Graham Kemp (telephone 772 54 11, room 6475 EDIT) The examiner will visit the exam room at 15:00 and 17:00.
Results:	Will be published by 26 March 2012 at the latest.
Exam review:	See course web page for time and place: http://www.cse.chalmers.se/edu/year/2012/course/MVE360/
Grades:	Grades for Chalmers students are normally determined as follows: ≥ 48 for grade 5; ≥ 36 for grade 4; ≥ 24 for grade 3. Grades for GU students are normally determined as follows: ≥ 42 for grade VG; ≥ 24 for grade G.
Help material:	None English language dictionaries are allowed.

Specific instructions:

- Please answer in English where possible. You may clarify your answers in Swedish if you are not confident you have expressed yourself correctly in English.
- Begin the answer to each question on a new page.
- Write clearly; unreadable = wrong!
- Fewer points are given for unnecessarily complicated solutions.
- Indicate clearly if you make any assumptions that are not given in the question.
- Write the page number and question number on every page.

Question 1. What output is printed when the following program is run?
4 p

```
#!/usr/bin/perl -w

$s = "gaatccaaat"; if ( $s =~ /c(.*)$/ )     { print "a) $1\n"; }
$s = "gaatccaaat"; if ( $s =~ /(a+)t/ )      { print "b) $1\n"; }
$s = "gaatccaaat"; if ( $s =~ /aa(.+)aa/ )   { print "c) $1\n"; }
$s = "gaatccaaat"; if ( $s =~ /(.)\1(.)\2/ ) { print "d) $1, $2\n"; }

$s = "HSLGSGSSTQAAS"; $s =~ s/S.//g     ; print "e) $s\n";
$s = "HSLGSGSSTQAAS"; $s =~ s/[GS]+/X/g ; print "f) $s\n";
$s = "HSLGSGSSTQAAS"; $s =~ s/^[^G]//   ; print "g) $s\n";
$s = "HSLGSGSSTQAAS"; $s =~ tr/GS/PT/   ; print "h) $s\n";
```

(4p)

Question 2. A seven-nucleotide motif, "TGGAAAG", occurs twice within a short exon that is conserved in different strains of the HIV-1 group M.
8 p

a) Write the code that needs to be added to the following Perl program so that the program reads data consisting of one exon sequence per line, and prints out those data lines that contain at least two occurrences of the motif TGGAAAG.

```
#!/usr/bin/perl -w

#
# Answer to part (a) goes here.
#

__DATA__
GGACAGCAGAGATCCACTTTGGAAAGGACCAGCAAAGCTCCTCTGGAAAG
GGACAGCAGAGATCCACTTTGCTATCGACCAGCAAAGCTCCTCTGGAAAG
GGACAGCAGAGATCCACTTTGGAAAGGACCAGCAAAGCTCCTCTGCTATC
GGACAGCAGAGATCCACTTTGGTAAAGACCAGCAAAGCTCCTCTGGAAAG
GGACAGCAGAGATCCACTTTGGAAAGGACTAGCAAAGCTCCTCTGGGAAG
GGACAGCAGAGATCCACTTTGGAAGGGACCAGCAAAGCTCCTCTGGAAAG
GGACAGCAGAGATCCACTTTGGAAAGGACTAGCAAAGCTCCTCTGGAAAG
```

(3p)

b) Write a Perl program that reads a single sequence from a FASTA format file whose name is specified on the command line, and counts how many "TGGAAAG" motifs that are present in that sequence. Here are the first few lines of a data file in FASTA format:

```
>gi|7452908|emb|AJ251057.1| Human immunodeficiency virus type 1
GGGTGCGAGAGCGTCAGTATTAAGTGGGGGAAAATTAGATGCATGGGAGAAAATTCGGTTAAGGCCAGGG
GGAAAGAAAAAATATAGACTAAAACATTTAGTATGGGCAAGCAGGGAGCTGGAAAGATTCGCACTTAACC
CTGGCCTTTTAGAATCAGCAGAAGGATGTCAACAACTAATGGAACAGTTACAATCAACTCTCAGGACAGG
```

(5p)

Question 3. The HindII restriction enzyme cuts DNA in the middle of the motif "GT[TC][AG]AC".
7 p

a) List all strings with 6 characters that match with the motif "GT[TC][AG]AC".
(1p)

b) Suppose that the Perl variable $s contains the sequence of a long piece of DNA. Write a piece of Perl code that "cuts" the DNA sequence in $s everywhere that the HindII restriction enzyme can cut, and put strings representing the resulting fragments of DNA sequence into array @frags.
(6p)

Question 4. 7 p

a) Assuming a match score of 2, a mismatch score of -1 and a gap score of -2, derive the score matrix for a global alignment of "CATAC" and "GTACC".
In this case, what is the score of an optimal global alignment? How many alignments have this optimal score (remember: each path represents a different alignment)?
(4p)

b) Draw a suffix tree for the string "ACACAC".
(3p)

Question 5. 4 p

a) We assume that you want to study the phylogenetic relationship between three different protein sequences, from cat, dog and man, respectively. Draw all the possible rooted and unrooted trees that are possible with these three sequences.
(1p)

b) Describe what is meant by a "distance based" method in molecular phylogeny.
(1p)

c) In a multiple alignment with four sequences from rat, mouse, human and chimpanzee you have one column (site) with this composition:

```
rat         A
human       G
mouse       A
chimpanzee  G
```

According to the principle of maximum parsimony, what is the most likely unrooted tree associated with this data? Draw the unrooted tree and explain why it is the tree preferred.
(2p)

Question 6. 14 p

a) What is the main advantage of next generation sequencing techniques compared to traditional capillary sequencing (Sanger sequencing)?
(3p)

b) Massively parallel pyrosequencing and Illumina sequencing are two next generation DNA sequencing techniques. Discuss the unique advantages of each technique in terms of sequencing performance. State also one preferred area of application for each technique.
(6p)

c) Describe the type and origin of the most common type of sequencing error encountered in massively parallel pyrosequencing.
(5p)

Question 7. 6 p

a) What is meant by the Markov property?
(1p)

b) What is the difference between a Markov chain and a hidden Markov model?
(1p)

c) What is the generalization in a Generalized HMM?
(1p)

d) What are the Viterbi, forward and forward-backward algorithms used for in HMMs?
(3p)

Question 8. 4 p

Assume that we have a sequence generating machine, consisting of only two states $S = \{A, C\}$, with initial probabilities $\pi = \{1, 0\}$.
Assume further that the machine generates a sequence of the form

AACAACAACAACAACAACAAC...

a) Is this a Markov chain? Show calculations.
(2p)

b) Is the 2nd order process a Markov chain? Show calculations.
(2p)

Question 9. 6 p

Assume we have three biased coins, A, B, and C, where the emission probabilities of the two sides of the coin, heads (H) and tails (T), are given by

	P(H)	P(T)
A	0.6	0.4
B	0.3	0.7
C	0.8	0.2

Assume that we start with either coin A or B with equal probabilities, then we switch between the two according to the outcome of C: if C shows 'H' we flip the same coin again, otherwise we change to the other coin. Now, assume that we have observed the sequence HHTH from the flipping of A and B.

a) What is the optimal state sequence of the model?
(4p)

b) Given the state sequence in (a), what is the probability of the observed sequence?
(2p)