CHALMERS UNIVERSITY OF TECHNOLOGY

Examination in Bioinformatics, MVE360

Monday 11 March 2013, 08:30-12:30

Examiner:	Graham Kemp (telephone 772 54 11, room 6475 EDIT) The examiner will visit the exam room at 09:30 and 11:30.
Results:	Will be published by 26 March 2013 at the latest.
Exam review:	See course web page for time and place: http://www.cse.chalmers.se/edu/year/2013/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:

- Check that you have received:
  - question paper (5 pages)
  - EMBL nucleotide sequence database entry SCYSP3G (1 page)
  - Perl reference guide (4 pages)
  - HMM formula collection (3 pages)
- 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 = "atccatattt"; if ( $s =~ /c(.*)$/ )         { print "a) $1\n"; }
$s = "atccatattt"; if ( $s =~ /(a+)t/ )          { print "b) $1\n"; }
$s = "atccatattt"; if ( $s =~ /.*(..)a(.*)..$/ ) { print "c) $1, $2\n"; }
$s = "atccatattt"; if ( $s =~ /(.)\1*$/ )        { print "d) $1\n"; }

$s = "atccatattt"; $s =~ s/a./g/g    ; print "e) $s\n";
$s = "atccatattt"; $s =~ s/.*at//g   ; print "f) $s\n";
$s = "atccatattt"; $s =~ s/t.*t/g/   ; print "g) $s\n";
$s = "atccatattt"; $s =~ tr/ta/cg/   ; print "h) $s\n";
```

(4p)

Question 2. 6 p

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

b) Write a Perl program that prompts the user to type in a string, reads that string, and then prints out all suffixes of that string.
(3p)

Question 3. 12 p

In an EMBL nucleotide sequence database entry, the positions where the coding sequence starts and ends are given on the feature table line (beginning "FT") for feature "CDS" (e.g. positions 908 and 2293 in EMBL entry SCYSP3G, see attached sheet).

a) Write a Perl program that reads the start and end positions of the coding sequence from an EMBL database file whose name is specified on the command line. The program should read the full sequence, and store the coding sequence in the variable $cs.
(4p)

b) Extend your solution to part (a) so that the program writes the coding sequence to standard output, with three nucleotides (one codon) per line.
(3p)

c) The codons "aca", "acc", "acg" and "act" all code for the amino acid threonine. Extend your solution to part (b) so that the program counts how often each of these four codons occurs in the coding sequence, and writes out the threonine codon(s) that occurs most often.
(5p)

Question 4. 4 p

Using a gap score of -2 and match/mismatch scores taken from the PAM250 substitution matrix (given below), derive the score matrix for a global alignment of "QFN" with "NGYE".

In this case, what is the score of an optimal global alignment? Give the alignment(s) with this score.

PAM250 substitution matrix:

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
A  2
R -2  6
N  0  0  2
D  0 -1  2  4
C -2 -4 -4 -5  4
Q  0  1  1  2 -5  4
E  0 -1  1  3 -5  2  4
G  1 -3  0  1 -3 -1  0  5
H -1  2  2  1 -3  3  1 -2  6
I -1 -2 -2 -2 -2 -2 -2 -3 -2  5
L -2 -3 -3 -4 -6 -2 -3 -4 -2  2  6
K -1  3  1  0 -5  1  0 -2  0 -2 -3  5
M -1  0 -2 -3 -5 -1 -2 -3 -2  2  4  0  6
F -4 -4 -4 -6 -4 -5 -5 -5 -2  1  2 -5  0  9
P  1  0 -1 -1 -3  0 -1 -1  0 -2 -3 -1 -2 -5  6
S  1  0  1  0  0 -1  0  1 -1 -1 -3  0 -2 -3  1  3
T  1 -1  0  0 -2 -1  0  0 -1  0 -2  0 -1 -2  0  1  3
W -6  2 -4 -7 -8 -5 -7 -7 -3 -5 -2 -3 -4  0 -6 -2 -5 17
Y -3 -4 -2 -4  0 -4 -4 -5  0 -1 -1 -4 -2  7 -5 -3 -3  0 10
V  0 -2 -2 -2 -2 -2 -2 -1 -2  4  2 -2  2 -1 -1 -1  0 -6 -2  4
```

(4p)

Question 5. 4 p

a) Draw all possible *rooted* trees with four different external nodes. (2p)

b) In the context of *maximum parsimony*, what is meant by an *informative site*? In the simple multiple alignment below, indicate the sites (columns in the alignment) that are informative.

```
GCTGT
GCTGG
ACACG
ACAGG
```

(2p)

Question 6. 6 p

a) Explain the property that characterizes a Markov Chain and make it a special case of a stochastic process.
(1p)

b) Explain the *processes* that compound a HMM. Which *parameters* describe the model? What is the difference between a HMM and a Markov Chain?
(3p)

c) Once you have trained a HMM, i.e. found the parameters, which algorithm would you use to parse a new DNA sequence? Briefly explain.
(1p)

d) How does a pair HMM differ from a standard HMM?
(1p)

Question 7. 4 p

Genes have variants called alleles. Each person has 2 copies (alleles) of a gene. Consider one gene with two alleles A and a in a population. So, in our supposed population we can have genotypes (possible combination of alleles for the same gene) $1 = AA$, $2 = aa$ and $3 = Aa$ with probabilities p_1, p_2 and p_3, respectively. We will consider our population is in the so called Hardy-Weinberg equilibrium, meaning that the genotype proportions remain the same from one generation to the next. For simplicity we will also consider that each individual has only one offspring. Now, starting with a randomly chosen individual in the population, at some point in time (generation 1), we let X_n = the genotype of her descendant in the nth generation.

a) Explain why $\{X_n, n = 1, 2, ...\}$ is a Markov Chain.
(1p)

b) What are the missing transitions, a_{13} and a_{22} in the transition matrix below?

$$\begin{bmatrix} p_1 + \frac{p_3}{2} & 0 & ? \\ 0 & ? & p_1 + \frac{p_3}{2} \\ \frac{p_1}{2} + \frac{p_3}{4} & \frac{p_2}{2} + \frac{p_3}{4} & \frac{p_1}{2} + \frac{p_2}{2} + \frac{p_3}{2} \end{bmatrix}$$

(2p)

c) If an individual has genotype Aa, what is the probability that her grandchild will have genotype Aa as well? Show how this probability can be calculated from the transition probabilities.
(1p)

Question 8. 4 p

Assume that we have a very simple HMM model that classifies stretches of genes in DNA sequences as exons (coding) or introns (non-coding). The emission probabilities are given in below:

	P(A)	P(C)	P(G)	P(T)
E	0.25	0.25	0.25	0.25
I	0.4	0.1	0.1	0.4

The transition probabilities between different states are 0.1, i.e
$a_{IE} = a_{EI} = 0.1$.
States are initiated with the same probability.

Find the optimal state sequence for the DNA stretch $TAAG$.

(4p)

Question 9. 2 p

Using a pair HMM it is possible to calculate the a score of similarity for two sequences Y_1^T and Z_1^U (where $Y_1^T = Y_1, ..., Y_T$ and $Z_1^T = Z_1, ..., Z_T$):

$$log \frac{P(Y_1^T, Z_1^U)}{P(Y_1^T, Z_1^U | R)} \tag{1}$$

The probability in the numerator is calculated under the assumption that the two sequences are generated by by the pair HMM, and the probability in the denominator is calculated under the assumption that they are randomly assembled, independently of each other.

a) By which algorithm can $P(Y_1^T, Z_1^U)$ be determined?
(1p)

b) In the context of sequence alignment, how is $P(Y_1^T, Z_1^U)$ to be interpreted?
(1p)

Question 10. 14 p

a) In metagenomics one sometimes talks about a *gene-centric analysis* approach. Describe what this means. (2-5 sentences)
(4p)

b) Describe and compare (giving advantages and disadvantages) two of the next generation sequencing techniques. Computational aspects should be included in your discussion.
(6p)

c) What is a *paired-end* sequencing technique, and how does a paired-end approach help in *de novo* genome assembly?
(4p)