Student Number

The University of Melbourne
Department of Computing and Information Systems
COMP 90016 Computational Genomics
Final examination: June 2016

Identical Examination papers: None.

Exam Duration: Three hours.

Reading Time: 15 minutes.

Length: This paper has 10 pages including this cover page and a blank page on the back of the cover page.

Authorised Materials: English language dictionaries and foreign language dictionaries are the only authorised materials. University regulations prohibit the use of electronic dictionaries.

Calculators: Calculators are not permitted.

Instructions to Invigilators: Students should be supplied with a script book. They may have additional script books on request.

Instructions to Students: There are 5 questions. You should attempt all questions. Answer all questions in your script book. Marks shown are out of 60. This paper counts for 60% of your final grade.
Write your student number on the top of this page.

This paper and any unused or discarded script books must be returned with your script book and must not be taken out of the examination room.

Library: The paper is to be lodged with the Baillieu Library.

Question 1: Coding for genes (Question 1 is worth 4 marks)

When genes are expressed in cells, a biological process first transcribes DNA to RNA and then translates RNA to proteins.
Since RNA gets translated to amino acids by triplets of nucleotides (codons), there are three distinct reading frames of the DNA/RNA: one that starts at the first base (with the first triplet comprising the first three bases), one that starts from the second, and one from the third base. Table 1 shows the translation of codons to amino acids.
Consider the sequence "AATGAACGTAGTATGTAGGCATTAAG".

a. Write down the amino acid sequence for *each* of the reading frames (ignoring partial codons at the start and end of the sequence). Use the three-letter acronym for each amino acid where applicable and comma separate the individual amino acids.
   (2 marks)
b. Which of the reading frames do you consider to be the most likely one to actually code for a protein and why?
   (2 marks)

Table 1: Translation tables from nucleotide codons (RNA) to amino acids.

Second base of codon

First base of codon	U	C	A	G	Third base of codon
U	UUU Phenylalanine phe UUC UUA Leucine leu UUG	UCU Serine ser UCC UCA UCG	UAU Tyrosine tyr UAC UAA STOP codon UAG	UGU Cysteine cys UGC UGA STOP codon UGG Tryptonphan trp	U C A G
C	CUU Leucine leu CUC CUA CUG	CCU Proline pro CCC CCA CCG	CAU Histidine his CAC CAA Glutamine gin CAG	CGU Arginine arg CGC CGA CGG	U C A G
A	AUU Isoleucine ile AUC AUA AUG Methionine met (start codon)	ACU Threonine thr ACC ACA ACG	AAU Asparagine asn AAC AAA Lysine lys AAG	AGU Serine ser AGC AGA Arginine arg AGG	U C A G
G	GUU Valine val GUC GUA GUG	GCU Alanine ala GCC GCA GCG	GAU Aspartic acid asp GAC GAA Glutamic acid glu GAG	GGU Glycine gly GGC GGA GGG	U C A G

Question 2: HMMs (Question 2 is worth 16 marks)

The lectures introduced Hidden Markov Models (HMMs) as a versatile problem solving technique in computational genomics. Consider the HMM shown in Figure 1 with six hidden states and transition probabilities between the states (only showing arrows for probabilities greater than 0). Table 2 shows the emission probabilities for each nucleotide (A, C, G, T) in each state of the HMM.

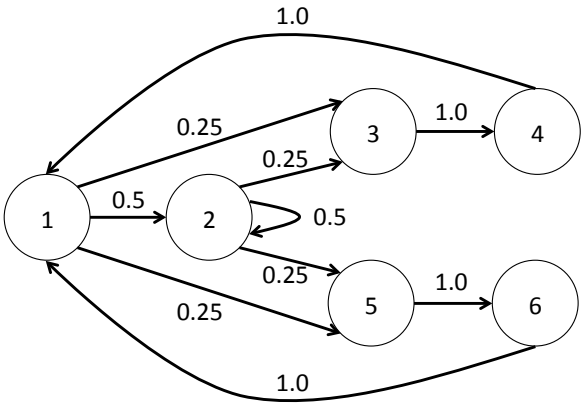

Figure 1: HMM with six states (state number in circle).

Table 2: Emission probabilities of HMM

State	A	C	G	T
1	1	0	0	0
2	0.25	0.25	0.25	0.25
3	1	0	0	0
4	0	0	1	0
5	0	0	1	0
6	0.1	0	0	0.9

The lectures also introduced the Viterbi algorithm for HMMs. The Viterbi algorithm uses the following recurrence equation for the dynamic programming matrix entry of state k and observation symbol i:

$$V_k(i) = e_k(i) \cdot \max_l (a(l,k) \cdot V_k(i-1)),$$

with $e_k(i)$ the emission probability of the ith observation in state k and $a(l,k)$ the transition probability of state l to state k.

Underflow is a problem for the Viterbi algorithm because products of probabilities can produce very small numbers. A way around this problem is a log transform of the probabilities. Since computing the log of a product of probabilities is the sum of the two probabilities' log, this simple trick allows a robust way to compute the Viterbi dynamic programming matrix. The recurrence relationship for log-transformed probabilities changes as follows:

$$V_k(i) = \log_2(e_k(i)) + \max_l (\log_2(a(l,k)) + V_k(i-1))$$

Table 3 shows the $\log_2$ values for each of the probabilities relevant to the HMM. Note that the maximum of two negative values selects the value closer to 0. For example, max(-2,-4) = -2.

Table 3: log2 transform of probability values relevant to the HMM. We define the logarithm of 0 as -∞ (minus infinity),

P	0	0.1	0.25	0.5	0.9	1
$\log_2(P)$	-∞	-3.3	-2	-1	-0.2	0

Questions:

a) What is the purpose of the Viterbi algorithm (in general)?
(2 marks)

b) Compute the Viterbi dynamic programming matrix for the given HMM and the observation sequence "AAGAGT". Use the log transformed technique outlined above, which allows you to compute sums of log values instead of products of probabilities. Note, taking the log of a probability equal to 0 is problematic. Table 3 defaults the result to negative infinity. Use this value whenever $P=0$ is part of the recurrence. For example, $-2 + -\infty = -\infty$.
Consider state 1 to be the starting state: Therefore, the first column of the matrix contains a 0 ($\log_2$ emission probability of "A" in state 1) for state 1 and $-\infty$ for all other states.
(8 marks)

c) What is the state with the highest value (log probability) for the last observation in the observation sequence (the "T")? What is the trace-back of states through all the observations from this maximum value? Write the sequence of observations and the sequence of according hidden states, so that they align according to your trace-back.
(4 marks)

d) Explain in a few sentences some purpose of the given HMM. That is, how could it be utilized in a genomics application?
(2 marks)

Question 3: CNVs and Clonality (Question 3 is worth 18 marks)

SNP arrays allow fast experimental evaluation of thousands of known SNPs in the human genome on a single chip. However, if large genomic rearrangements have occurred in the sample DNA, these too can be detected from the chip data from a combination of absolute and relative intensities of the probe pairs.
Observe the plots of log-ratio and B-allele frequencies of a population of tumor cells from a patient in Figure 2. Assume an overall ploidy of two in the data – i.e., the average intensity (logR=0) relates to having two copies of a chromosome.
To aid your calculations below, Table 4 shows the log2 values of data points of interest.

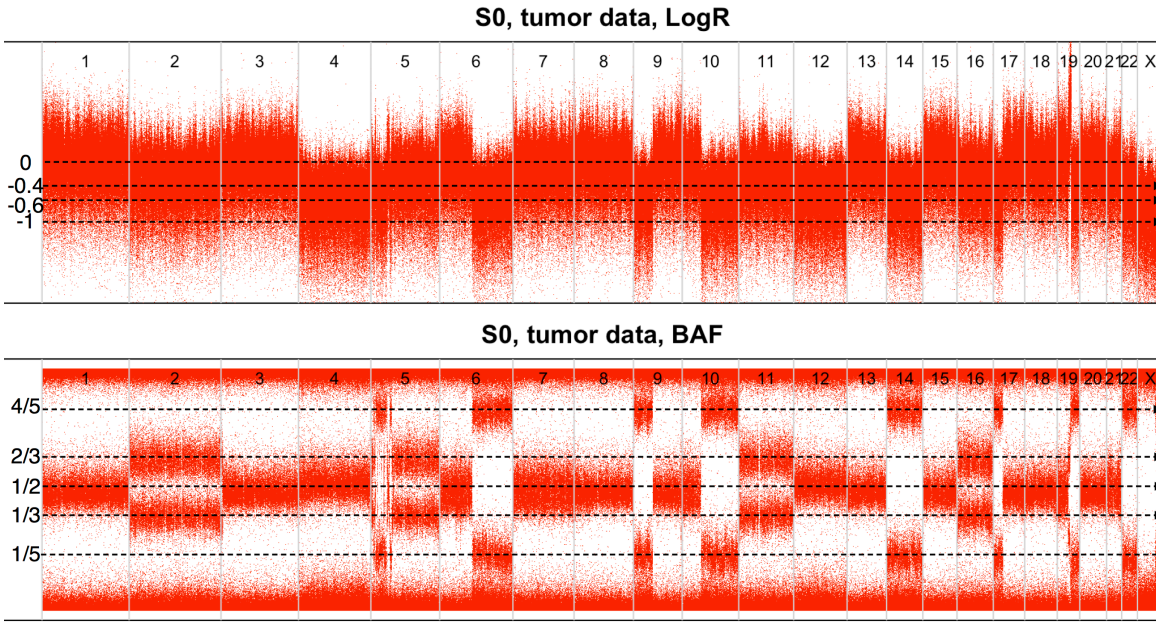

Figure 2: SNP array data of a population of tumor cells. The chromosomes are ordered along the x-axis of the plots. The top panel shows the log2 ratio of the probe intensities compared to the average intensity. The bottom panel shows the B-allele-frequency (BAF). For your convenience lines with key log values and frequencies are shown as dashed lines through the panels. Note that the y-axis tick marks are not quite to scale, but allow for easier calculations.

x	0.5	0.625	0.75	1
log2(x)	-1	-0.68	-0.42	0

Table 4: selected log2 values for values relevant to the question. Approximate the readings in Figure 2 with the closest values.

$$\log R = \log 2\left(\frac{((vA+vB)c+2-2c)}{2}\right), BAF = \frac{vB\cdot c+1-c}{(vA+vB)c+2-2c}$$

Equation 1: Formula to calculate logR and BAF for a given allele, when vA and vB reflect the copy number of the paternal and maternal alleles and c the clonality of vA and vB within the population of cells.

Questions:

a) Averaging over the entire array, do the overall copy numbers for chromosomes tend to go up, down or stay neutral? Back up your answer with a few observations from the data.
(2 marks)

b) Is the sampled DNA from a male or a female patient? Discuss both logR and BAF in your answer.
(2 marks)

c) What kind of event can produce a signal, such as that on chr6 or chr9. You do not have to compute the copy number event numerically; instead argue how the signal pattern can arise.
(2 marks)

d) The BAF panel shows a recurring pattern of frequency across all the chromosomes. It is reasonable to assume that copy number events with the same BAF are in fact from the same clone within the population of cells. Therefore we can identify two major clones within the data that have suffered copy number changes to their DNA. For your convenience lines in the logR and BAF panels identify data points with similar values.
As a starting point, assume the copy number changes in chr2 have occurred in clone A and the changes in chr14 in clone B. Equation 1 may help you with the following questions.
    1. For all cells likely to be in clone A, which chromosomes have undergone copy number changes?
    2. For all cells likely to be in clone B, which chromosomes have undergone copy number changes?
    3. For all cells in clone A, what is the copy number change and what is the clonality of clone A? Back up your answer with data points in Figure 2 and show any calculations necessary to arrive at your conclusion.
    4. For all cells in clone B, what is the copy number change and what is the clonality of clone B? Back up your answer with data points in Figure 2 and show any calculations necessary to arrive at your conclusion.
    5. Your answers to 3. and 4. above gave you clonality values for clone A and B (if you could not derive values from the data, assume any two different clonal proportions other than 0 or 100% and continue with these). All cells in clone A share the same copy number changes as do all cells in clone B. Draw a schematic of cells and colour or circle parts of them to roughly (or exactly) represent the two clones. How can the number of cells that are affected by either clone be minimized?

(10 marks)

e) Observe and discuss the signal of the logR and BAF panels for chromosome 4 in Figure 2 and anything that is peculiar about it. What kind of event(s) could lead to this signal? How does your answer reflect on the number of cells affected by copy number changes as discussed in d)5?
(2 marks)

Question 4: SNPs (Question 4 is worth 6 marks)

In the Lectures and Assignment 2 we discussed a publication on the eye color of humans based on particular SNPs in the HERC2 gene. Table 5 shows the genotypes of test subjects in the study and their eye color. You have learned that parents pass on one of their copies of each chromosome to their offspring at random. Assume a person (let us refer to her as "Maxx") with genotype A/A, C/C (each on the reverse strand) for the rs1129038 and rs12913832 SNPs, respectively.

	rs1129038	rs12913832	rs916977	rs1667394	Phased[a]	Blue	Brown
1[b]	A/A	C/C	G/G	A/A	**AC**GA/**AC**GA	169	1
2[b,c,d]	A/G	C/T	G/G	A/A	**AC**GA/GTGA	1	47
3[c,d]	A/G	C/T	A/G	A/G	**AC**GA/GTAG	10	81
4[d]	A/G	C/T	G/G	A/G	**AC**GA/GTGG	0	8
5	G/G	T/T	A/A	G/G	GTAG/GTAG	0	18
6	G/G	T/T	A/G	A/G	GTAG/GTGA	0	14
7	A/A	C/C	A/G	A/G	**AC**GA/**AC**AG	1	1
8	A/G	C/C	A/G	A/G	**AC**GA/GCAG	1	1
9	A/G	C/C	G/G	A/A	**AC**GA/GCGA	1	1
10	G/G	T/T	A/G	G/G	GTAG/GTGG	0	1
11	G/G	T/T	G/G	A/A	GTGA/GTGA	0	3

[a] **AC** phased rs1129038 and 12913832 in bold.
[b] The Fisher exact test comparing counts for row 1 versus 2, $p = 2 \times 10^{-16}$.
[c] Row 2 versus 3, $p = 0.097$.
[d] Row 2 versus 3 versus 4, $p = 0.195$.

Table 5: Table from the publication by Sturm et al. showing the relationship of SNPs in the HERC2 gene and human eye color.

Questions:

a) What is the most likely eye color of Maxx? What is the probability of this to be the true eye color (only based on the results in Table 5)?
(2 marks)

b) What possible haplotypes could each of the parents have?
(2 marks)

c) Which of the potential parents' haplotypes do not occur in Table 5? Argue why not all of the haplotypes occur in the table.
(2 marks)

Question 5: Structural Variants (Question 5 is worth 16 marks)

a) Name at least one structural variation that is copy number neutral and one that is copy number sensitive.
(2 marks)

b) Comparing structural variant detection with anomalous paired-end reads and with read depth, name one benefit that paired-end reads offer over read depth in regards to their ability to detect variants.
(2 marks)

c) A sequencing experiment produced paired-end reads from a human DNA sample. The read pairs are sequenced from fragments with an average length of 400nt. The standard deviation of the fragment size distribution is 50nt. After alignment, read pair A is mapped to the same chromosome with the first read on the forward and the second read on the reverse strand. The outer distance between the two reads (5' to 5') is 460nt. Read pair B is mapped in similar fashion, but the distance between the two reads is 210nt. Answer the following questions:
    1. Is the mapping of read pair A indicative of a structural variation? And if so, what is the type of variation, its approximate size, and approximate location?
    2. Is the mapping of read pair B indicative of a structural variation? And if so, what is the type of variation, its approximate size, and approximate location?

    (4 marks)

d) Figure 3 presents a schematic of read pairs mapping to a reference genome. The figure also shows a read depth track at the top, which shows qualitative changes in sequencing coverage (assume reads other than the ones in P1-P4 to contribute to the read depth signal).
Read pairs P1-P4 align discordantly to the reference sequence. Discuss:
    1) For each read pair, what is the most basic type of structural variation that they indicate?
    2) Considering all four read pairs at the same time, as well as the read depth track, what is the complex rearrangement indicated by the data (use the breakpoint indicators "A", "B", "C" to refer to genomic locations on "ChrA")?
    3) Assuming that the genotype for the rearrangement above is heterozygous, draw a schematic of both haplotypes of the donor chromosome pair ChrA according to your solution determined in d) 1).

    (8 marks)

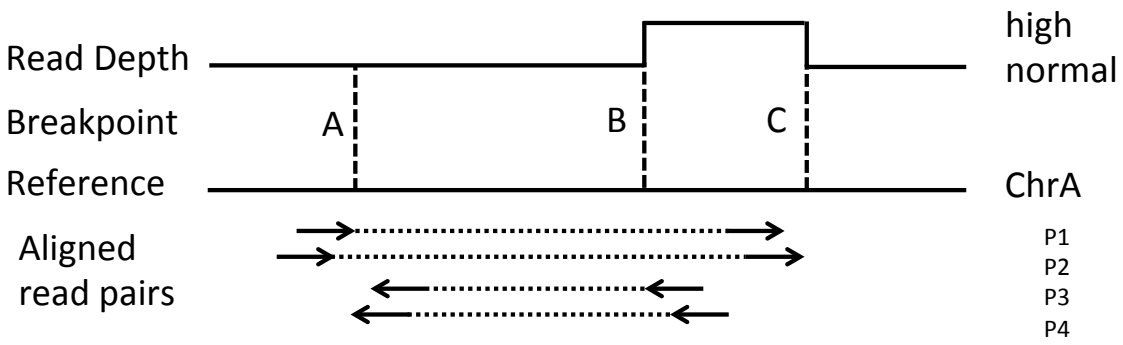

Figure 3: Schematic of a sequencing experiment. On the top the read depth is shown qualitatively ("high"/"low") across a reference genome (called "ChrA"). Three breakpoints ("A", "B", "C") are marked along the reference. Beneath four example read pairs ("P1" to "P4") are shown: Reads are shown as solid black arrows (the direction of the arrow indicating the strand that the read aligns to on ChrA); dashed lines between reads indicating that they are from the same read pair.