Examination in MVE510 - Introduction to bioinformatics
Time: Wednesday August 28 2019, 08.30-12.30
Examiner: Erik Kristiansson, phone 070-5259751
On call: Olof Zetterqvist, phone 031-7725325
Aids: None

Maximum is 40 points. Grade 3 requires 18 points, grade 4 requires 26 points and grade 5 requires 34 points.

1. Next generation DNA sequencing.

   (a) Describe the main differences between traditional DNA sequencing (e.g. Sanger sequencing) and next generation DNA sequencing (e.g. Illumina sequencing). (3p)

   (b) Give two examples of application areas where next generation sequencing has had a major impact. Describe also why. (3p)

2. Let the scoring matrix S be defined as

$$S(a,b) = \begin{cases} 7 & \text{if } a = b \\ -3 & \text{otherwise} \end{cases}$$

   (a) Use a linear gap penalty with $d = -5$ to calculate the score of the alignment

```
Sequence 1: AGTCCGGA-
Sequence 2: -GAC-GCAT
```

   (2p)

   (b) Use an affine gap penalty with $d = -5$ and $e = -3$ to calculate the score of the alignment

```
Sequence 1: GGCTTTGCA
Sequence 2: GCGC---CA
```

   (3p)

3. In a research project, scientists are investigating germline mutations in wheat associated with tolerance to cold. Next generation sequencing is used to sequence the genome at a coverage of 20 times.

   (a) Describe the main steps in bioinformatics analysis of finding ('calling') mutations in a genome. (3p)

   (b) The scientists aims to extend the project and they have two alternatives: 1) increase the coverage or 2) include more biological replicates (i.e. wheat samples). Discuss the advantage of both approaches. What would you recommend? (3p)

4. In an RNA-seq experiments, the expression of human genes are investigated in liver tissue from five patients. The abundance of each gene is quantified by mapping the RNA-seq reads to a reference and counting the number matches for each gene. The result for two different genes is shown in the table below together with the total number of reads for each patient. The lengths of the genes are show in the last column.

Gene	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Gene length
Gene 1	1000	1500	250	150	100	1000
Gene 2	5000	8000	1200	12000	8000	2000
Total	5×10^6	1×10^7	1×10^6	3×10^6	2×10^6	

   (a) Describe the normalization method RPKM (reads per kilobase per million). In what way does it reduce the variability in the data? (2p)

   (b) Use RPKM to calculate normalized values for the genes in the table above. (2p)

   (c) Assume that samples 1,2 and 3 are from healthy individuals and samples 4 and 5 are from sick individuals. Calculate the average fold-change between these two groups. Interpret the results. Are the genes up or down-regulated? (2p)

5. In bioinformatics, a large number of tests are often performed simultaneously. This gives rise to a multiple testing problem.

   (a) Describe the multiple testing problem and what it can lead to if not properly addressed (2p)

   (b) Describe the Benjamini-Hochberg method of false discovery rate (FDR). How can it be used to solve the multiple testing problem (3p)

   (c) Assume that we search for differentially expressed genes in a transcriptomic dataset. The dataset contains in total 20,000 genes. The five most significant genes have p-values 1.5×10^{-6}, 4×10^{-6}, 2.5×10^{-5}, 6×10^{-5} and 8×10^{-5}. Use the Benjamini-Hochberg method to correct the p-values. What is the false discovery rate if the significance threshold is set after the fifth gene? Explain the result. (2p)

6. Assume that the nucleotide sequence AGCAGAGC$ represent a part of a bacterial gene ($ is the termination character).

   (a) Calculate the suffix tree for the sequence (3p).

   (b) Demonstrate how you can use the suffix tree to find the subsequence AGA (3p).

7. In metagenomics, genes are often quantified by counting reads. This leads to data that does not follow the normal distribution. Describe at least two statistical approaches that does not rely on normal assumptions and that can be used to identify differentially abundant genes in metagenomic count data. (4p)

GOOD LUCK!!