Examination in MVE510 - Introduction to bioinformatics
Time: Friday April 6 2018, 14.00-18.00
Examiner: Erik Kristiansson
On call: Erik Kristiansson, phone 070-5259751
Aids: None

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

1. In modern genomics, next generation sequencing platforms are used to read DNA molecules.

   (a) Explain the main differences between traditional Sanger sequencing technology and the next generation sequencing technology. (3p)

   (b) Describe the difference between short and long reads. Name one sequencing platform that produces short reads and one that produces long reads. Name one advantage and one disadvantage with long reads. (3p)

2. Use the Smith-Waterman algorithm to perform a local alignment between

   ```
   Sequence 1: ACGTAT
   Sequence 2: CTTA
   ```

   Use the scoring matrix

   $$S(a,b) = \begin{cases} 5 & \text{if } a = b \\ -4 & \text{otherwise} \end{cases}$$

   and a linear gap with the penalty $d = -3$. Report all alignments and their corresponding scores. (5p)

3. In a research study, gene expression analysis is performed on colon cancer biopsies to identify differentially expressed genes involved in tumor development. The study contains two groups of patients, one with patients that have malignant tumors and one with patients that have benign tumors. The gene expression is measured for each gene ($n = 25,000$ genes) using RNA-seq and the resulting data is normalized and log-transformed. A linear model is used to identify the differentially expressed genes.

(a) Formulate a linear model that can be used to identify genes that are differentially expressed between the groups. Introduce suitable variable names and model assumptions. (3p)

(b) In the study, the age and gender of each patient is documented. Extend the linear model in a) by including these variables. Why may the extended model be better for finding differentially expressed genes? (2p)

(c) What statistical assumptions are done in the analysis in a) and b)? Are these assumptions true? What may happen if the assumptions are violated? (2p)

4. Normalization is an important step in the identification of differential expressed genes in data from transcriptomics.

   (a) What is normalization? What is it used for and why is it important? (3p)

   (b) The normalization methods 'counts per million reads' (CPM) and 'reads per kilobase per million' (RPKM) are based on the total counts in a sample. State at least one reason why this can introduce errors into the data. (2p)

5. In many areas of bioinformatics, such as resequencing, transcriptomics and metagenomics, thousands of genes or mutations are statistically analyzed simultaneously. This is known as multiple testing.

   (a) What are the main statistical issues associated with multiple testing? (2p)

   (b) Describe the Bonferroni method for correction of p-values. How does it address the issues in a)? (2p)

   (c) Assume that you search for differentially expressed genes in a transcriptomic dataset. The dataset contains in total 10,000 genes. The three most significant gene have the p-values 3.75×10^{-7}, 7.32×10^{-7} and 2.12×10^{-6}. What are the corresponding Bonferroni-corrected p-values? How many genes would be significant if the cut-off for the corrected p-values were 0.01? (2p)

6. Assume that the nucleotide sequence AGTAGC$ represent a part of a human gene (where $ is the termination character).

   (a) Calculate the Burrows-Wheeler transform (BWT) of the sequence. (3p)

   (b) Use the results from a) to demonstrate how the occurrence of the sequence GTAG can be found in AGTAGC. (3p)

7. Resistance to the fluoroquinolone antibiotics is caused by mutations in the gyrase A gene (gyrA). In a research study, researchers aimed to investigate the effect that traveling has on the abundance of mutations in gyrA in the human gut microbiome. The hypothesis was that traveling to countries where antibiotic bactera are more prevalent may increase the abundance of mutated gyrA genes in the gut microbiome.

   (a) How would you design the study given that you have unlimited resources? Describe the choice of study design, which subjects to include, what samples to collect, sequencing methodology to use, and the principal steps in the data analysis. (3p)

   (b) Which are the main complications in analyzing mutations in a microbiome? How do these complications differ compared to analyzing mutations in the genome of a single organism? (2p)

GOOD LUCK!!