Examination in MVE510 - Introduction to bioinformatics
Time: Thursday January 17 2019, 08.30-12.30
Examiner: Erik Kristiansson
On call: Henrik Imberg, phone 5325
Aids: None

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

1. Data produced by next generation DNA sequencing platforms typically needs to be pre-processed.

   (a) What does pre-processing mean and why is it necessary? (2p)

   (b) Name one method that is used for pre-processing of DNA sequence data. Describe how it works and what it aims to achieve. (2p)

2. Apply the Needleman-Wunsch algorithm to perform a global alignment between

   ```
   Sequence 1: TATTT
   Sequence 2: TACTTG
   ```

   Use the scoring matrix

   $$S(a,b) = \begin{cases} 7 & \text{if } a = b \\ -5 & \text{otherwise} \end{cases}$$

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

3. In a scientific study, whole genome sequencing was used to identify single nucleotide polymorphisms (SNPs) associated with inherited human diseases.

   (a) What is the difference between germline and somatic mutations? Which type would be of main interest in the study? (2p)

   (b) Describe the main bioinformatical steps necessary to identify SNPs from sequence reads generated by next generation DNA sequencing. (3p)

4. In gene-centric analysis of shotgun metagenomic data, the abundance of genes are compared between samples from different experimental conditions.

(a) Describe briefly the bioinformatical process of gene quantification ('binning'). What is the difference between direct and guided binning? (3p)

(b) Name at least two statistical approaches that can be used to identify differentially abundant genes. What are their strengths and weaknesses? (3p)

5. Unsupervised analysis of data.

   (a) Describe the difference between supervised and unsupervised data analysis. Describe a scenario where unsupervised data analysis would be of interest. (2p)

   (b) Describe the main difference between hierarchal clustering and k-means clustering. (3p)

6. The Burrows-Wheeler transform (BWT) is often used for comparing sequence reads against a reference.

   (a) What properties of the BWT makes it attractive for finding exact matching of sequencing reads against a reference? (2p)

   (b) Calculate the Burrows-Wheeler transform (BWT) for the sequence ATCGAT$ ($ is the termination character). (3p)

   (c) Demonstrate how the original sequence can be reconstructed from the BWT calculated in (b). (2p)

7. RNA-seq is sometimes used to find mutations. The advantage is that this enables both analysis of gene expression and identification of various DNA alterations from the same data.

   (a) How would you design an RNA-seq experiments which aims both to analyze gene expression and to identify mutations? For example, what sequencing platforms would you recommend and what aspects needs to be taken into account when deciding the sequencing depth? (2p)

   (b) Describe briefly the analysis steps that you think are necessary to identify mutations from RNA-seq data. (3p)

   (c) Discuss what forms of mutations that you expect to find using this approach. Are there any mutations that may be missed in the RNA-seq data that could have been identified from whole genome sequencing data? (3p)