Examination in MVE510 - Introduction to bioinformatics
Time: Friday April 26 2019, 14.00-18.00
Examiner: Erik Kristiansson
On call: Fanny Berglund, 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. Long read sequencing platforms that are able to read DNA fragments of tens of thousands base pairs have recently become available.

   (a) Summarize the advantages and disadvantages of using long reads generated by current sequencing platforms in relation to using short reads. Describe also at least one application where long reads are preferred over short reads. (3p)

   (b) Select one long read sequencing platform and describe how it works. (3p)

2. Use the Needleman-Wunch algorithm to perform a global alignment between

   ```
   Sequence 1: AGTCCG
   Sequence 2: TCTAG
   ```

   Use the scoring matrix

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

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

3. In a scientific study transcriptomics was used to investigate the gene expression patterns in breast cancer. Two groups of patients were included in the study, one group consisting of patients diagnosed with cancer (biopsies were taken from the tumor) and one control group consisting of healthy patients (biopsies were taken from normal tissue). The gene expression was measured for each gene ($n = 20,000$) using RNA-seq and the resulting data was normalized and log-transformed.

(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. (2p)

(b) How many parameters are needed to be estimated in order to fit the model to all of the genes in the dataset? (1p)

(c) Assume that each patient group can be further divided into the categories 'old' (older than 50 years) and 'young' (younger than 50 years). Update the model in (a) to include this new information. (1p)

(d) Compare the model in (a) and (c). What are the advantages of the model in (c)? Can you see any advantages of the model in (a)? (2p)

4. In metagenomics, sequencing of the 16s rRNA gene is used to estimate the taxonomic diversity in a sample.

   (a) What is a 'hypervariable' region and why are they important when parts of the 16s rRNA gene are sequenced using short read platforms? (2p)

   (b) Provide an overview of the bioinformatical analysis necessary to estimate the taxonomic diversity from 16s amplicon data. What steps are essential? (3p)

5. Normalization is an important step in the analysis of transcriptomic and metagenomic data.

   (a) What is the purpose of normalization and why does it improve the interpretation of the data? (2p)

   (b) Describe the different steps of the normalization method 'trimmed mean of M-values' (TMM). (3p)

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

   (a) Calculate the suffix tree for the sequence (2p).

   (b) Demonstrate how you can use the suffix tree to find the subsequence TAG (2p).

   (c) What is the main disadvantage of using suffix trees to mapping of reads to the human genome? (2p)

7. In shotgun metagenomics, DNA fragments are randomly selected from microorganisms and then sequenced.

   (a) Discuss shotgun metagenomics in relation to amplicon metagenomics. What are the benefits of using shotgun metagenomics? What are the main challenges? (3p)

   (b) When designing an shotgun metagenomics experiments, the number of samples that are going to be sequenced and the sequencing depth per sample needs to be decided. If we assume that we have a fixed budget for the experiment, a large number of included samples will result in a lower sequencing depth and, vice versa, a higher sequencing depth will limit the number of samples that can be included. Discuss these parameters and their importance. What could happen if an experiment includes too few samples? What could happen if a too low sequencing depth is used? (4p)

GOOD LUCK!!