Examination in MVE510 - Introduction to bioinformatics
Time: Thursday January 16 2020, 08.30-12.30
Examiner: Erik Kristiansson
On call: Felix Held, phone 031-772 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 from next-generation DNA sequencing is known to contain a wide range of errors and inaccuracies.

   (a) Describe the meaning of the following types of errors: 'substitution', 'indel' and 'duplicate'? (2p)

   (b) What is 'adapter contamination'? Why is this an issue and where does it come from? (2p)

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

   ```
   Sequence 1: GTCTAGT
   Sequence 2: CTCAG
   ```

   Use the scoring matrix

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

   and a linear gap penalty with $d = -8$. Report all optimal alignments and their corresponding scores. (5p)

3. In metagenomic amplicon sequencing, a specific genomic region ('barcode') is sequenced and analyzed to assess the taxonomic composition of a microbial community.

   (a) Provide an overview of the analysis of data from amplicon sequencing. Which are the necessary steps and why are they important? (3p)

   (b) One way to analyze amplicon sequence data is to estimate the diversity. What is the difference between richness and evenness and what do they measure? Describe one way to estimate richness and one way to estimate evenness (3p)

4. In bioinformatics, a large number of tests is often performed simultaneously. This gives rise to the multiple testing problem.

   (a) Describe the multiple testing problem. What can happen if it is not properly addressed? (2p)

   (b) What is the family-wise error rate (FWER)? Also, describe the Bonferroni method and how it can be used to control the FWER (3p).

   (c) Assume that we search for mutations in a genome consisting of 10,000,000 positions. The five most significant genes have the p-values 7×10^{-12}, 2×10^{-11}, 3×10^{-9}, 3×10^{-8} and 7×10^{-7}. Use the Bonferroni method to adjust the p-values. How many positions are significant if we require the FWER to be less than 0.01? (2p)

5. Shotgun metagenomics is used to study bacterial communities by sequencing random fragments of their collective genome (metagenome).

   (a) What does binning mean? What is the difference between direct and reference-guided binning? Which approach would you recommend if you have very short sequence reads? Why?(3p)

   (b) Why do we normalize data from shotgun metagenomics? Also, mention at least one type of systematic error that may be present in metagenomic data but is not present in transcriptomic data. (3p)

6. Unsupervised methods are often used to find patterns in high-dimensional data.

   (a) What is the main difference between clustering and principal component analysis (PCA)? Give examples on when to use clustering and when to use PCA. (3p)

   (b) How does hierarchical cluster analysis work? Describe the different steps of the algorithm and explain meaning of a 'distance metric' and a 'linkage criterion'. (3p)

7. In RNA-seq, RNA is converted to cDNA which is then sequenced. This results in reads that are randomly selected from the RNA molecules in the sample.

   (a) Briefly describe how reads from RNA-seq can be used to estimate gene expression (i.e. RNA abundance). (2p)

   (b) Explain the difference between quantifying the expression of genes, exons and isoforms. (2p)

   (c) Data from RNA-seq is often said to be semiquantitative. Explain what this means and its implications, i.e. what can we not measure using RNA-seq because of that the data is semiquantitative? (2p)

GOOD LUCK!!