Facit to 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. Note that the answers to the questions are partial and many thus not be sufficient for full score.

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)
   Answer: Substitutions means that a nucleotide has be incorrectly substituted by another nucleotide. An indel means that a nucleotide has either been deleted or inserted. A duplicate is multiple sequencing of the same physical DNA molecule that gives rise to identical reads. See lecture notes for full details.

   (b) What is 'adapter contamination'? Why is this an issue and where does it come from? (2p)
   Answer: Adapter contamination is when parts of the adapter are still present in the reads. Removal of the first 5' adapter is straight-forward. When the DNA fragments are very short, there is a risk that we also read parts of the 3' adapeter. This is much hard to remove and may result in adapter contamination. See lecture notes for full details.

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)
   Answer: Two alignments, each with a score of 20.

```
Alignment 1:
CT-AG
CTCAG

Alignment 2:
TCTAG
TC-AG
```

3. In 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)
   Answer: The main steps are pre-processing, OTU clustering, OTU annotation and Statistical analysis. See lecture notes for full details.

   (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)
   Answer: Richness estimates the number of species present in a community while evenness measures how even their abundance are. Richness can be estimated by counting the number of unique OTUs in sample. Evenness can be estimated by Shannons or Simpsons index. See lecture notes for full details.

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)
   Answer: The multiple testing problem arise when testing many hypotheses. If we, for example, test 100 tests, which at a level of 0.05 we will get 5 significant test even when all null hypotheses are true. The challange with multiple testing is thus to differentiate between false positives and true positives. See lecture notes for full details.

   (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).
Answer: FWER is the probability that at least one of the tests results in false positives. The Bonferroni method adjusts the p-values by multiplying them with the number of performed tests. This will control the FWER. See lecture notes for full details.

(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)
Answer: After adjustments, i.e. multiplication by 10,000,000, two tests are significant (i.e. have p-values¡0.01).

5. Shotgun metagenomics is used to study bacterial communities by sequence 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? (3p)
   Answer: Direct binning tries to identify genes directly from the reads. Reference-guided binning maps the reads to a reference, which has been annotated for the presence of genes. If you have short reads, the reference-guided binning should be used since it is very hard to identify genes directly from the reads. See lecture notes for full details.

   (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)
   Answer: Data from shotgun metagenomics is affected by a wide range of systematic errors and needs to be normalized before samples can be compared. One type of systematic errors that is present in metagenomic data and not transcriptomic data is differences in the taxonomic composition where e.g. the proportion of bacteria can differ significantly between samples. See lecture notes for full details.

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)
Answer: Clustering is the process where we group samples into clusters based on their similarity. PCA is a method to visualize high-dimensional data in lower dimensions by picking dimensions that describe most of the variability. Clustering is thus used when we want to group samples while PCA is used to explore patterns in the data. See lecture notes for full details.

(b) How does hierarchical cluster analysis work? Describe the different steps of the algorithm and the explain meaning of a 'distance metric' and a 'linkage criterion'. (3p)
Answer: Hierarchical clustering works by successively merge samples into larger and larger clusters. The distance metric defines how to measure the distance between two samples. The linkage criterion defines how to measure the distance between two clusters. The algorithm works by merging the closest clusters until all samples are in one single clusters. The results are typically visualized in a dendrogram describing at what distances the different clusters merged. See lecture notes for full details.

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

(a) Briefly describe how reads from RNA-seq can be used to estimate gene expression (i.e. RNA abundance). (2p)
Answer: Gene expression is estimated by mapping reads to a reference, typically the full genome of the investigated the organism. By counting the number of matches to each gene, we can estimate its expression. See lecture notes for full details.

(b) Explain the difference between quantifying the expression of genes, exons and isoforms. (2p)
Answer: Quantification of genes means that we count all reads matching to the gene. Quantification of exons means that we count all reads matching to the individual exons. Quantification of isoforms means that we try to estimate the abundance of each possible gene product (isoform). See lecture notes for full details.

(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)
Answer: The data from RNA-seq is semiquantitative in the sense that it is quantitative but only in relation to other genes. This

means that we do no have information about the total number of RNA molecules in the sample. Thus, RNA-seq data can not distinguish between an increase of expression in one gene and the decrease in expression of all other genes. See lecture notes for full details.

GOOD LUCK!!