HOME EXAM

FSK-2053: DATA SCIENCE AND BIOINFORMATICS

FOR FISHERIES AND AQUACULTURE

INSTRUCTIONS: Select **two questions of your choice** out of the three in the Data Science part and **two questions of your choice** out of the three in the Bioinformatics part. All questions have the same score. Maximum score is 80 marks. The answers have to be provided in a single PDF file. You can use MS-Word or Libre-Office to edit this file. Include all R scripts you have written to answer the questions as text boxes in this file. Plots can be included as figures. You can also use screenshots to illustrate the answers of the Bioinformatics part. This is an individual exam, so discussion is **not** allowed.

The exam will be presented on the 31st of May and the answers have to be delivered before 18th of June at 1400.

Data Science Question 1 (20 marks total)

Files needed:

Question1_dataset.csv

Question1_metadata.csv

The table Question1_dataset.csv contains metabarcoding read counts for Metazoans detected in environmental DNA from water samples collected along the coast of Norway. Each line represents a MOTU (a proxy for species) and includes the taxonomic information of that MOTU, and a series of columns with read count abundances in the samples. Sample names are coded by the characters "FDIR" and a 3-digit number. Three replicates have been analysed for each sample, which are coded using "A", "B", or "C" after the sample name. We are interested in investigating the metabarcoding read counts of detected fishes in this dataset using the *tidyverse* package.

1. Open the file in R Studio. Using the *tidyverse*, select only the rows belonging to fishes (classes "Actinopterygii" or "Chondrichthyes") and save the results in another table. How many MOTUs of fishes are there in the set (2 marks).

2. Transform the dataset of fishes to tidy format. You can select the read count data columns whose name start with "FDIR" for selecting the columns with the data that needs to be put in longer format. The names of the columns will go to a new column called "sample_rep" and the values will go to a new column called "reads". How many observations has the tidy table? What is the total sum of the reads from fishes in that table? (2 marks).

3. We will collapse the three replicates of each sample. To do this, separate the sample_rep column into two new columns containing the sample name and the replicate name. You can call the new columns "sample" and "replicate". Now group this table by scientific_name and sample to generate a new collapsed table by summing the reads for each species from the three replicates of each sample. The new table should have three columns (scientific_name, sample, and sum_reads). How many observations has this collapsed table? (2 marks).

4. Table Question1_metadata.csv contains information for the samples. Join the collapsed table from the previous step with the metadata table, by sample name. Keep only those samples that exist in both tables. Count the number of unique samples in the metadata table and the fish reads table, and compare it with the joint table. How many samples without metadata have been lost? (4 marks)

5. We will collapse now the total reads for each scientific_name (species) by geographic area. Generate a new table with the total reads for each species in each of the five areas. How many observations has this table? (2 marks). We want to order the names of the areas geographically. Use the command ordered() to define the Area

column as an ordered factor with this order: ("Ishav", "Nordland", "Midt", "Vest", "Sør"). This is the order that will be displayed in the graphics.

6. Using geom_boxplot, generate a boxplot to show the distribution of the total number of reads for each fish species by Area (integrating all species). Use logarithmic transformation in the reads axis. The Areas must be shown in the right order ("Ishav", "Nordland", "Midt", "Vest", "Sør"). (2 marks)

7. From the dataset of collapsed reads by area, select only those rows that have counts >0. Then group by species and generate a new table by calculating the total number of areas in which each fish species is present. This new table should have just two columns: fish species and number of areas in which these species are present (1 to 5). How many rows has this table? (2 marks)

8. From the dataset of collapsed reads by area, select those fish species that have been detected in just 1 area. We will call these "endemic fishes". Generate a barplot showing the logarithm of reads vs the five geographical areas, faceted by species (with free vertical scales) for these endemic fishes. This plot should allow you to easily count how many "endemic fishes" are there in each area. Which area has the most detections of "endemic" fish species and how many such species have been detected there? (3 marks).

9. From the dataset of collapsed reads by area, select only those fishes that have been detected in either 4 or 5 areas. Those species are the "widely distributed" ones. Generate a barplot showing the logarithm of reads vs the five geographical areas, faceted by species (with free vertical scales) with these widely distributed fishes. How many widely distributed fishes are there in our dataset? Three of these fishes have read abundances that are clearly decreasing from North to South. Which are their names? (3 marks).

Note: the R script performing all the asked operations and analyses has to be provided in text format.

Data Science Question 2 (20 marks total)

Files needed:

Question2_table.csv

Question2_metadata.csv

The file Question2_table.csv contains the abundances (counts) of six groups (phyla) of algae in several normalized samples along the coast of Norway. Sample identifiers are coded by the characters "FDIR" and a 3-digit number. The file Question2_metadata.csv contains information about the samples, including the month, the latitude and the longitude.

1. Open file Question2_table.csv in R and transform it to tidy format. How many observations are there in the tidy table? (2 marks).

2. Join the tidy-formated dataset to the metadata file, keeping only the observations that are in both files. The names of the columns will go to a new column called "sample" and the values will go to a new column called "counts". How many observations are there in the joint table? (2 marks).

3. Generate a plot with points representing algal counts vs latitude, grouped and coloured by month and faceted by phylum_name (set the scales free). Generate a similar plot, but representing the counts vs longitude. (2 marks).

4. Select the observations for Ochrophyta and save them in a new database. Build a linear model of the counts of Ochrophyta as a function of Month, Latitude and Longitude, modeling the residuals as gaussian. Include all interactions in the model, besides the main effects. If the interactions are not significant, then repeat the model excluding the interaction terms. Which variable(s) have a significant effect in the counts of Ochrophyta? Give the p-values for the significant effects. You can ignore the assumption checking of linear models. (3 marks).

5. Select the observations for Cryptophyta and save them in a new database. Build a linear model of the counts of Cryptophyta as a function of Month, Latitude and Longitude, modeling the residuals as gaussian. Include all the interactions in the model, besides the main effects. If the interactions are not significant, then repeat the model excluding the interaction terms. Which variable(s) have a significant effect in the counts of Cryptophyta? Give the p-values for the significant effects. You can ignore the assumption checking of linear models. (3 marks).

6. Select the observations for Bacillariophyta and save them in a new database. Build a linear model of the counts of Bacillariophyta as a function of Month, Latitude and Longitude, modeling the residuals as gaussian. Include all the interactions in the model, besides the main effects. If the interaction of month with any of the numerical variables is significant, then repeat the analysis separately by creating a different model for each value of the month. Which variable has a significant effect on the counts of Bacillariophyta, and for which months? Give the p-values for the significant effects, obtained from the separated analysis. You can ignore the assumption checking of linear models. (4 marks).

7. Select the observations for Haptophyta and save them in a new database. Build a linear model of the counts of Haptophyta as a function of Month, Latitude and Longitude. Include all the interactions in the model, besides the main effects. Which main effects appear to have significant effect in this model? If the interaction of month with any of the numerical variables is significant, then repeat the analysis separately by creating a model for each value of the month. Which variables have a significant effect on the counts of Haptophyta, and for which months? Give the p-values for the significant effects, obtained from the separated analysis. You can ignore the assumption checking of linear models. (4 marks).

Note: the R script performing all the asked operations and analyses has to be provided in text format.

Data Science Question 3 (20 marks total)

Files needed:

Question3_data.csv

Question3_herring_landings.csv

The file Question3_data.csv includes data of fish eggs and larval abundances from North Atlantic surveys by ICES during the 1950-2021 period. Columns included are: year, month, stage, survey, species and quantity.

1. Load the dataset in *tidyverse*. Summarise the total number of specimens found for each species. Plot the total number of specimens (logarithm-transformed) for each species in a horizontal barplot in *ggplot*. Remove the legend for a better visualization, How many different species are present in this dataset? Which species has the highest value of detected larvae+eggs? (2 marks).

2. Create a new table containing only the most common species. To do this, calculate the number of observations for each species and select only those species with more than 30 observations in the dataset. How many species are there in this reduced table? (2 marks).

3. In order to study the seasonality of the fish and larvae, filter the original table to select only common species (with more than 30 observations). Then create a table summarising by Species and Month. Plot a vertical barplot of total abundances vs month, faceted by species. In the view of these results, name four species which have their reproductive season in winter. (4 marks).

4. In order to explore the long-term abundance trends over the years, filter the original table to select only common species (with more than 30 observations). Then create a table summarising the total sum of individuals found by species and year. Draw vertical barplots of total abundance vs year, faceted by species. In the view of the plots, and asumming that the sampling efforts have remained unchanged since the surveys started (for each species), name three species which have experienced significant declines in their larval abundances over the years. (4 marks).

5. Do you find any strange pattern in the yearly abundances of five of the most abundant species? Hint: Select all observations whose data come from the Survey named GORL, arrange the data by year and month and look at their abundances. Propose a possible explanation which could have caused these errors. (4 marks).

6. After removing all observations coming from the GORL survey in the original dataset, select all observations of herring. Calculate the total numbers of herring eggs+larvae for every year and save these values in a new table. The file question3_herring.pdf contains the data of yearly landings of herring in the state of Maine. Load this file and join the two tables by year, keeping only the years appearing in both tables. Produce a scatter plot of total herring larvae+eggs vs total pounds of herring landed in Maine. Add a linear trend line. Calculate a linear model to test the correlation between these two variables. Is there a significant correlation between them? Write a possible explanation of why this relation has a negative slope. (4 marks).

Bioinformatics Question 4 (20 marks total)

All the files required for following part are placed in the folder /net/common in the server fsk-2053.azure.uit.no

File needed: Q4_final.fasta

Huge demand and unsustainable fishing practices have led to the decline of shark populations across the globe. Sharks play an important role in structuring the marine ecosystem and food webs. Sharks are commercially important for their meat and fins. Late maturation, low fecundity and longevity made them vulnerable to overexploitation. Continuing unsustainable fishing has now triggered the immediate conservation actions by FAO and CITES. Stock assessment have been hampered by lack of species-specific data. In shark fishing countries, landings are often officially recorded using generic categories which implies different classes of shape and size with no mention of species. In this case, researchers used a molecular technique called DNA barcoding to identify the shark meat collected in the local markets to species level, so that responsible authorities can use this information for management of shark fishery in Taiwan.

1. Define DNA barcoding. What is the standard DNA marker used in this technique? Explain the principle behind the DNA barcoding? And provide references wherever it is necessary. (4 marks)

2. The sequence data for this question is in the folder: /net/common/question_4, which is placed in the folder: /net/common/ in the server. You need to copy that sequence file from there and place it in your home directory. All the analysis should be performed at home directory with appropriate folder name. The modules and databases are placed in /net/common.

a) Identify the unknown sequences generated from sequencing of DNA extracted from shark meats collected at various local markets in Taiwan using locally run blast search engine. Present the code used to run the local blast search engine and explain the code in words. Also provide the statistical evidence used to make the choice (8 marks).

How many shark species are recovered from these sequences? Remember multiple sequences may represent one species. Present the results in a table form (4 marks).

b) What are the possible explanations if two databases provide different identity for a query sequence? (2 marks)

c) Provide the conservation status of these shark species using a appropriate database. Explain the results from the stock management perspective. (2 marks)

Bioinformatics Question 5 (20 marks total)

Files needed:

1. Q5_final_query.fasta

2. Q5_reference.fasta

3. Q5_sample_labelling.txt

Increasing consumer demand for seafood combined with concern over the health of our oceans has led to many initiatives aimed at challenging destructive fisheries and promoting sustainable fisheries. An important global threat to sustainable fisheries originates from the illegal, unreported, and unregulated (IUU) fishing and associated seafood frauds. Current estimates suggest that around 25% of fishery catches come from IUU practices. To tackle this problem, EU enacted a new law obliging the inclusion of species names on catch labels throughout the distribution chain. Such certification measures do not, however, guarantee accuracy of species designation. For the dataset that you will use for Question 5, researchers collected the fish flesh samples from different retails in a quest to investigate potential mislabeling of gadids. All the files required for this question are placed in the server /github page.

a) What is the right designation of these species? Use BOLD systems. Use Q2_sample_labelling.txt and Q2_final_query.fasta to solve this question. Q2_sample_labelling.txt contains the sample code, species reported on the packaging. Fasta file contains the sequence for those corresponding mislabeled samples. Identify the probable correct species for these mislabeled samples (6 marks).

Also provide the next closest species and support your finding by providing the similarity scores (2 marks).

b) Another way to identify unknown samples/sequences is through phylogenetic trees.

What is phylogenetic tree? Explain the steps to construct a phylogenetic tree in general. Construct a neighbour joining tree with statistical support to test the conclusion of BOLD systems. Mention which substitution model you used and what statistical test has been done to test the significance of branching. Refer Q2_reference.fasta for the sequences of reference species. Q2_final_query.fasta consists of sequences you need to identify. Compare the results from phylogenetic tree and BOLD systems briefly (9 marks).

c) What are the applications of DNA barcoding? Explain briefly. Provide references wherever it is necessary (3 marks)

Bioinformatics Question 6 (20 marks total)

File needed:

1. Question6_query_sequence.fasta

A biological database is a large, organised body of persistent biological data usually associated with a software designed to update, query, and retrieve the data stored within the system. Please refer to the Question6_query_sequence.fasta file located in the folder: /net/common/ in the server. This file contains the nucleotide sequence required to solve this question.

a) Explain the "central dogma of molecular biology". How is it related to the structure and functioning of a biological database? Provide references wherever it is necessary (4 marks)

b) Identify the sequence. Extract the exact location of this sequence in the Atlantic salmon genome using any appropriate database. Briefly explain the steps. Specify the gene Id and name of the gene. Partition the hit into different parts of a gene and show the structure of the gene. (8 marks)

c) What's the role of this gene in Atlantic salmon biology in general. Why it is important to understand this gene from the farming of Atlantic salmon point of view? Provide references (8 marks)