Computer Exercise 4:
Metagenomic data analysis

Introduction to Bioinformatics (MVE510)
Autumn, 2023

Introduction

It is time for the fourth and final computer exercise in which we will focus on the analysis of metagenomic data. In this exercise, we will work with data from amplicon sequencing of the 16s marker gene as well as shotgun sequencing of all DNA in a sample containing a complex mixture of microorganisms. Your role in this exercise is to serve as a bioinformatician at a center for environmental monitoring in Mexico. In 2010, the BP drilling platform 'Deepwater Horizon' located in the Mexican Gulf exploded and sank, resulting in large quantities of crude oil being released into the aquatic ecosystem. Crude oil contains polycyclic aromatic hydrocarbons (PAHs) which are known to be highly toxic to a wide range of organisms. There is therefore a concern that the oil exposure has resulted in changes in the ecosystems, in particular in the microbial communities living in the sediments on the seabed. You have therefore been given the important task of investigating the environmental effects of oil exposure.

Figure 1: The Deepwater Horizon was an offshore drilling rig that exploded and sank 2010 causing massive releases of oil affecting the aquatic ecosystems.

In order to investigate the effect of the oil exposure, you ordered samples to be collected both close to and far away from the borehole. The latter samples will serve as a control in this study. You also arrange sequencing of the samples, both based on the amplicons from the 16s marker gene and the total amount of DNA in the sample (i.e. shotgun metagenomics). It is now your task to make sense of the resulting data. You are particularly interested in any differences in biodiversity but also changes in the taxonomic composition and the biochemical functions as a result of the oil exposure.

The data that you will use in this computer exercise has already been pre-processed and the first steps in the bioinformatics analysis have been done. For the amplicon data from the 16s rRNA gene, this means that the data has been clustered to form operation taxonomic units (OTUs). These OTUs have then been annotated taxonomically. The abundance of each OTU has been estimated in each sample by counting the number of matching reads. The shotgun

metagenomic data has been analyzed by binning the reads against a reference which has been functionally annotated using the TIGRFAM database. This database describes a wide range of bacterial genes and is described at https://www.jcvi.org/tigrfams. The abundance of each TIGRFAM gene has been estimated by counting the number of matching reads. The data and the corresponding annotation are available in four files *16s_counts.txt*, *16s_annotation.txt*, *gene_counts.txt,* and *gene_annotation.txt* which can retrieved from http://bioinformatics.math.chalmers.se/courses/MVE510/.

Similar to previous exercises, all steps of the analysis should be done in R. The computer exercise should be performed in groups of a maximum of two students and the results will be examined through a written report that describes the different steps you took, the generated figures, and your conclusions. The written code should be added as an appendix. The reports should be handed in through the course home page in Canvas, latest January 18th. As always, don't forget to read and follow the specific guidelines available on the home page regarding the structure and formats in which the report should be handed in.

Exercise 4.1

We will start with analyzing the data from the amplicons of 16s rRNA data. Use **read.table** to load the files 1*6s_counts.txt* and 16s*_annotation.txt* into R. Note that both of these files are of the exact same length and that each row in each of the files corresponds to one single OTU. Examine the data and describe what you see. How many counts in total do the different samples have? How is the annotation file structured? Why do you think that the annotation is incomplete for some of the OTUs?

Important: The function **read.table** has a special interpretation of certain characters that can appear in annotation files. Examples of such characters are **"**, **'** and **#** where **"** and **'** are interpreted as quoting characters while **#** is interpreted as comment characters. In order to read the annotation properly, this functionality needs to be turned off. We need also to specify that the file is tab-separated and contains a header. This means that the following arguments need to be added to **read.table**: **comment.char="", quote="", sep="\t", and header="TRUE".**

Several of the OTUs have very few counts and do therefore not provide much information about their abundance. Remove these OTUs by filtering the data. A suitable cut-off is to require each OTU to have at least 5 reads in total over all samples.

Exercise 4.2

Similarly to transcriptomic data, unsupervised methods can be used to analyze the overall structure of the metagenomic data. Use an unsupervised method of your choice (e.g. clustering or PCA) to investigate how the six samples are structured. Is there a separation between samples from high and low oil exposure?

The data we are working with in this exercise are counts, which have different statistical properties than e.g. continuous data from a normal distribution. One such property is that the variance is dependent on the expected value, which higher counts have an overall higher variability (remember the Poisson distribution which has the same expected value and variance). A variance stabilizing transformation can be used to remove, or at least reduce, this dependence which makes data easier to interpret. Apply a variance-stabilizing transformation in the form of

$$f(x) = log(x + 1)$$

and redo the analysis above (unsupervised analysis using a method of your choice). Can you see any difference? Did the samples within the groups become more or less homogenous?

Exercise 4.3

Next, we will estimate and compare the diversity of the samples. However, most ways to calculate the diversity are dependent on the number of reads in a sample. For example, the number of unique species detected in a sample increases with the number of reads – a high sequencing depth thus leads to more detected species. In order to make samples comparable we need to make their sequencing depth equal. This is typically done by a process called rarefaction. When a sample is rarefied, reads are randomly selected *without replacement* until a pre-specified number of reads has been reached. The number of reads is typically set to a number less than the number of reads in the sample with the lowest sequencing depth.

Implement a function that takes as input three arguments: 1) a character vector of OTU name/identifier, 2) a vector of counts for the OTUs, and 3) and number indicating the resulting sequencing depth. The function should produce a rarified sample to the specified sequencing depth and return the corresponding OTUs and their counts. Show that the function works by rarifying the 16s count data. Set the sequencing depth to a suitable number.

Hint: Given a character vector of OTUs and a numeric vector with their counts, the following lines can be useful:

```
reads<-rep(OTUs, times=counts)
reads.sample<-sample(reads, size=10000, replace=FALSE)
counts.sample<-as.data.frame(table(reads.sample))
```

In this code, **OTUs** a character vectors with the OTU names and **counts** a vector of the same length but with the counts of each OTU. Note that the code has pre-specified the number reads to 10,000 – change this so it fits your data. The last line of the code is necessary to ensure that the result is a data.frame (and not a 'table-type' object that is produced by the function **table**). When you run this code, check the results from each line so that you are aware of how it works.

Exercise 4.4

The diversity of microbial community measures reflects the present species and their abundance distribution. There are several ways to measure diversity and, in this computer exercise, we will work with *richness* and *evenness*. Richness describes how many unique species are present and can be estimated by calculating the number of OTUs with at least one read. The evenness describes, in contrast, the uniformity of the species distribution, i.e. if the abundance of different species is similar or if some species dominate. For a sample, the evenness can be estimated by calculating the so-called Shannon index H' which is defined as

$$H' = -\sum_{i=1}^{N} p_i \log(p_i)\,.$$

Here N is the total number of OTUs and p_i is the relative abundance of OTU i.

Implement functions to estimate the richness (unique number of species) and evenness (Shannon's index). Apply the functions to the data rarified by the function you implemented in Exercise 3. Describe the results. Do you see any difference in diversity between samples from high and low oil exposure?

Exercise 4.5

Next, we turn the attention to specific OTUs. We are especially interested in differentially abundant OTUs, i.e. OTUs with an altered abundance due to the oil exposure. In contrast to computer exercise 3, where we used a linear model on transformed data, we will here use methods that are specially developed for count data. Can you find any arguments why it may be especially important to work directly with the count data in this exercise?

The method that we will use for the data is called DESeq2. 'DE' in DESeq stands for differential expression and the method was originally developed for data from transcriptomics (RNA-seq). It has, however, been shown that it also works very well for metagenomic data. DESeq2 uses a statistical model that models the specific structure of count data generated when counting fragments from DNA sequencing. This means that it often has a higher power than models assuming a Gaussian distribution, especially when there are few counts and few samples. Please see the lecture notes for more details about DESeq2.

DESeq2 should already be installed on the Chalmers computers. If you are using your own computer, you need to install DESeq2 yourself. For more information, see https://bioconductor.org/packages/release/bioc/html/DESeq2.html. Once installed, DESeq2 is loaded into R using the **library** command.

To apply DESeq2, we need to specify which of the samples are exposed to high and low concentrations of oil, respectively. This is done using a so-called design matrix. In this particular case, where we compare two groups, the design matrix can be specified using a **data.frame** consisting of a single column, where samples exposed to high levels of oil are indicated with '1' while samples exposed to low levels of oil are indicated with '0'. This tells DESeq2 that we are interested in comparing the samples exposed to high levels of oil to those that are exposed to low levels of oil where the latter is set as a reference. We can thus create the design matrix using

design.matrix<-data.frame(exposure=c(1,1,1,0,0,0))

Note that the order of the 0s and 1s needs to match the samples specified in the *16s_counts.txt* file.

Once the design matrix has been defined, we can run DESeq2. This consists of two steps. First, the data and the design matrix need to be combined into a 'dataset', which is the form of data object that DESeq2 is using. This can be done using the command **DESeqDataSetFromMatrix**, i.e.

counts.ds<-DESeqDataSetFromMatrix(countData=counts, design.matrix, design=~exposure)

The last argument is R formula that tells DESeq that we want to compare the samples based on exposure. Note that we do not need to use the rarefied data since the statistical model implemented in DESeq can properly handle data with different sequencing depths.

The next step is to apply the model and test each OTU for differential abundance. This can be done using

res.ds<-DESeq(counts.ds)

Finally, DESeq2 has a function called 'results' that can be used to print out a list of the results. Note that this list is ordered in the same way as the counts you have supplied. Note also that you need to specify the arguments **independentFiltering=FALSE** and **cooksCutoff=FALSE** in order to ensure that adjusted p-values are provided for all OTUs.

Apply DESeq to identify differentially abundant OTUs between samples that are exposed to high and low concentrations of oil. Combine the results with the annotation and sort the result based on the p-value. How do you interpret the adjusted p-value? Set a reasonable significant cut-off and describe how many OTUs are significant.

Exercise 4.6

Examine the ten most significant OTUs from the analysis done in exercise 5. Describe their taxonomy. Do these bacteria increase or decrease in the oil-contaminated samples?

Previously, bacteria from the families Alteromonadaceae and Thiotrichales have been shown to be able to degrade the polycyclic aromatic hydrocarbon (PAH) present in crude (see the abstract of https://www.ncbi.nlm.nih.gov/pubmed/22709320 but note that these bacteria are named by their genera Alteromonas and Cycloclasticus). Are bacteria from these families present in your result? Do they increase or decrease in the exposed sediments?

Exercise 4.7

We will now turn our attention to the shotgun metagenomic data. Similarly to the 16s data, this information also consists of counts. However, in this case, the counts come from binning reads based on their function. In this particular dataset, the gene counts have been classified based on the TIGRFAM database, which describes a wide range of bacterial genes and functions. To become familiar with this database, go to http://tigrfams.jcvi.org/cgi-bin/index.cgi, select a few TIGRFAM terms, and examine their annotation.

Due to the similarity between 16s and gene count data, they are often analyzed using almost the same approaches. This is also something that we will do in this computer exercise. Use **read.table** to read *gene_counts.txt* and *gene_annotation.txt* into R. Make sure that you understand the structure of the data and the annotation. How many reads do you have for each sample?

Use a filter to remove genes with very few counts similar to what you did in Exercise 1. Repeat the steps in Exercise 2 and use an unsupervised method of your choice to explore the data. Do the samples separate according to the level of exposure? If not, discuss why this may be the case.

Exercise 4.8

Analyze the diversity in the sample by calculating the richness and evenness. Use the functions you implemented in Exercise 4. Note that the data, as before, needs to be rarefied in order to make the results comparable between samples. What do richness and evenness mean when it comes to gene count data? Do you see any differences between the samples?

Exercise 4.9

Use DESeq2 to identify differentially abundant genes between samples that are exposed to high and low concentrations of oil. How many genes are significant? Are the relative abundance of the most significant genes increasing or decreasing?

There are several bacterial genes and pathways that are hypothesized to be involved in oil degradation. This includes various forms of dehydrogenases, in particular the gene pdxA (4-hydroxythreonine-4-phosphate dehydrogenase).

Find the gene pdxA in the gene list. Does it increase or decrease in the contaminated samples?

Hint: **grep** can be used to search for a pattern in a vector of strings.

Exercise 4.10

Summarize and discuss the effects of oil exposure based on all results in the computer exercise. In what way does the exposure seem to affect the bacterial communities?