Culture-independent Methods for Studying Microbial Communities

Contents

Introduction

This document provides background information on microbial ecology and a short tutorial in a bioinformatic application for this field. Also included are questions to help frame your understanding of the subject and data generated from the tutorial, as well as a page including links for further learning and practice.

The materials for this course are located here:

https://github.com/jmicrobe/BI331-taxonomy

Learning goals

Students will learn how DNA sequenced-based, culture-independent techniques are used to study microbial communities. Students will formulate hypotheses and test them using bioinformatic and statistical tools to compare microbial communities.

Acknowledgments

This exercise was originally written and graciously provided by Dr. Ann Klein in the Institute of Ecology and Evolution at the University of Oregon. It has been updated and modified using new analysis techniques by Jessica Hardwicke of the Masters of Bioinformatics program at UO.

Background

This section covers relevant background information for understanding what microbial ecology is, how bioinformatics can be used to understand microbial data from the environment, and how these elements relate to the data we will be using for the exercise.

What is microbial ecology?

What is a microbial community? Ecological communities are associations of species that co-occur in the same location at the same time. Microbial community ecology is a field of scientific study focused on the biotic and abiotic factors that determine how populations of microorganisms associate to form communities, how communities interact with each other, and how communities interact with the environment[1].

What is bioinformatics?

Bioinformatics is a field of science that combines aspects of biology, computer science, and information technology to analyze biological information using computers and statistical techniques. The primary goal of bioinformatics is to develop software tools to generate useful biological knowledge. In microbial ecology, bioinformatic tools are typically used to analyze DNA sequence data. These data can include sequences from cultured organisms or sequences from communities. The latter is discussed in more detail below.

Why use DNA sequencing to study microbes?

Greater than 99% of the microbes found in nature cannot be cultured and studied in the laboratory. Techniques have been rapidly advancing in recent years making it possible to study microbes and their communities in new ways without the need for culturing. To study these "unculturable" organisms, scientists study their DNA (and sometimes RNA). They collect a sample like a gram of soil or a liter of seawater and extract all the DNA in that sample. Most commonly, a specific gene in that DNA, the 16S rRNA gene, is amplified using polymerase chain reaction (PCR). This produces many copies of all the 16S rRNA genes present in the sample. The 16S genes are sequenced using "high-throughput" DNA sequencing technology that generates thousands (and sometime up to millions) of 16S gene sequences present in every sample. From these 16S gene sequences, we can learn which microbes are present in the sample and their relative abundance.

The 16S rRNA gene is one gene of choice for studying microbial communities for three primary reasons. First, all bacteria have at least one copy of the gene in their genome. Second, there are regions of the gene sequence that are highly variable (which allows us to distinguish among closely related bacteria) and other regions that are more conserved (which allows us to compare among more distantly related groups). Third, it is phylogenetically conserved. This means that closely related bacteria have more similar 16S rRNA gene sequences than do distantly related bacteria.

What is "Open Science"?

The scientific data and tools used in this exercise are made possible because of a push for what's known as "open science". Open science allows for greater collaboration, reproducibility, and accountability with regards to scientific data. The OpenScience project lists the following fundamental goals of open science:

- Transparency in experimental methodology, observation, and collection of data.
- Public availability and reusability of scientific data.
- Public accessibility and transparency of scientific communication.
- Using web-based tools to facilitate scientific collaboration.

For our purposes, open science provides an excellent avenue for learning about how bioinformatics tools are used without having to generate our own microbiome data.

Exercise data background

Study data

For this exercise you will be analyzing 16S rRNA gene sequences from human gut samples. The samples were collected from children in Europe (14 samples) and a rural African village in Burkina Faso (14 samples). The research for this study was published by Filippo et al. (2010)[5] and the data made publicly available on the European Nucleotide Archive website.

The human gut microbiome

Microorganisms in the human gastrointestinal tract (GI) perform functions and produce compounds essential for our health. They break down complex molecules (e.g. complex polysaccharides) from the food we eat making them easier for our cells to digest. They produce essential vitamins (e.g. B12 and K) and amino acids, which are absorbed by cells in the gut and then used

throughout the body[1]. This community of microorganisms inhabiting our gut is known as the human gut microbiome.

The composition and function of the gut microbiome varies among individuals and populations. Factors such as ethnicity, diet (e.g. diets rich in protein versus diets rich in complex carbohydrates), and genetics are known to contribute to the variation we see in the human gut microbiome[2]. Variation in the composition of these microbial communities also leads to variation in their function. One interesting example of this is a study that compared the gut microbiome of Japanese and American populations[3]. Researchers found a high abundance of bacteria carrying a specific gene in Japanese populations compared to American populations. This gene helps the microbes break down complex carbohydrates found in the seaweed used to make nori, one of the main ingredients in sushi and a common ingredient in other Japanese cuisine. Because seaweed is a larger part of the diet in Japanese populations compared to American populations, the guts of Japanese populations harbor this seaweed-digesting bacterium, which helps them get the most energy possible from their food. (There's another interesting level to this story. The gut bacteria that carry this gene actually acquired it though horizontal gene transfer from marine bacteria that colonize the seaweed in the ocean!). This is just one of many examples of how diet can affect the composition and function of gut microbial communities.

Exercise overview

You will perform two analyses using tools developed by the Ribosomal Database Project[4]. In the first, you will assign taxonomic classifications to the 16S rRNA gene sequences in the data set you will be provided using the Naïve Bayesian Classifier. This tool classifies the sequences in your data by comparing them to 16S rRNA gene sequences from well-characterized bacterial species using an algorithm based on Bayesian probability. The Naïve Bayesian Classifier generates a hierarchical classification (Domain, Phylum, Class Order, Family, Genus, and Species) for each 16S gene sequence in your data.

For the second analysis, you will statistically compare the composition of two communities (BF versus EU). The Library Compare Tool tests for significant differences between the two communities in the relative abundances of taxonomic groups. It uses Bayesian probability to estimate the likelihood that the relative abundance of a given taxonomic group is the same for the two communities. The output from this analysis includes taxonomic groups for which the relative abundance is significantly different, and the significance value of the difference. The significance value can be thought of as the likelihood that the difference is due to chance. The lower the significance value, the more confident we can be that the difference in relative abundance reflects real differences in how the communities are structured.

References

1. Madigan, Michael T., et al. “Biology of Microorganisms.” (2012).
2. Yatsunenko, Tanya, et al. “Human gut microbiome viewed across age and geography.” Nature 486.7402 (2012): 222-227.
3. Hehemann, Jan-Hendrik, et al. “Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota.” Nature (2010): 908-912.
4. Cole, J. R., et al. “The Ribosomal Database Project: improved alignments and new tools for rRNA analysis.” Nucleic Acids Research (2009): D141-D145.
5. De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., … Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proceedings of the National Academy of Sciences of the United States of America, 107(33), 14691–6.

Analysis

This tutorial will work you through the steps needed to complete the assignment. You should not need to install anything if you are using the laptops provided by the lab, but if you choose to do this on your own computer you'll need the following (totally free) programs:

- Git
- RDP Tools
- RStudio

Note that the RDP classifier is also available online, however the output file may be slightly different from the command-line generated file we will use in class.

Cloning the class repository to your computer and previewing the files

To make things simple we will work from the desktop. Open terminal and enter:

```
cd ~/Desktop
```

Now we need to clone the class repository, which will copy the files from GitHub and create a folder in the current directory:

```
git clone
   https://github.com/jmicrobe/BI331-taxonomy.git
```

Change to the new directory, using `cd`:

```
cd BI331-taxonomy
```

Check that all the files are there by using the `ls` command to list what's in the folder:

```
ls
```

You should see the following:

```
ANALYSIS.md      BF_samples.fasta    LINKS.md
rscript.R   ASSIGNMENT.md        EU_samples.fasta
README.md        taxonomy_guide.pdf
```

Let's peek into the .fasta files. The fasta format is used to contain sequencing data. We can use the head command to see the first several lines of the file:

```
head BF_samples.fasta
```

Your output should have repeating lines that look something like this:

```
>10BF.397096_4143 F4HTPAO04ECMBR
   orig_bc=TCAGATCAGACAC new_bc=TCAGATCAGACAC
   bc_diffs=0
GTCCGCACGGTAAACGATGGATGCCCGCCTGTTGGTCTGAATAGGTCAGCGGCC
AAGCGAAAGCATTAAGCATCCCACCTGGGGAGTACGCCGGCAACGGTGAAACTC
AAAGGAATTGACGGGGGCCCGCACAAGCGGAGGAACATGTGGTTTAATTCGATG
ATACGCGAGGAACCTTACCCGGGCTTGAATTGCAGAGGAAGGATTTGGAGACAA
TGACGCCCTTCGGGGTCGTCTGTGAAGGTG
```

The above represents **one entry** in the fasta file. In the context of this study this represents one individual sequence of a 16S rRNA gene from a human fecal sample. How many individual entries do we have in this particular file?

You could count these entries manually, but we're lazy programmers so we want to find a way to make the computer do the work. Note that each entry has two lines: The first line, known as the header, starts with a > followed by what looks to be metadata. The second line is the corresponding nucleotide sequence, `GTCCG...` etc. (more info on the fasta format).

With bash (the language you've been using in the terminal) there are many utilities to explore files, including the `ls` and `head` commands we used earlier. Let's try `grep`, which can be used to search a file and return matching queries. Since each entry starts with > we can look for that:

```
grep '>' BF_samples.fasta
```

Whoa! That printed out a bunch of lines, all starting with >. What `grep` did was find and return **all** of the header lines in the fasta file.

But that still doesn't tell us **how many** entries there are, and we're still too lazy to count them all. We can use the utility `wc` which stands for "word count" and the option `-l` which will only count the number of lines. But we want to do that on the output of our `grep` search, not the entire file, so we will use a `|` (pipe) to do both at once:

```
grep '>' BF_samples.fasta | wc -l
```

You should have gotten `1190` as output. This means that in our BF_samples.fasta file there are 1190 entries, or 1190 16s rRNA sequences.

Taxonomic classification of sequences

Even though we counted the number of entries in our BF_samples.fasta file to be 1190, that doesn't mean there are 1190 *different* species in our sample. As mentioned in the background, we can use the RDPTools classifier to assign taxonomy information to our sequences.

Enter the following command in terminal, noting that `/path/to/classifier.jar` will need to be changed depending on the location of `classifier.jar` on your computer.

```
java -Xmx1g -jar /path/to/classifier.jar -o
   BF_classified.txt -h BF_hier.txt BF_samples.fasta
```

Repeat the previous command, but for EU_samples.fasta:

```
java -Xmx1g -jar /path/to/classifier.jar -o
   EU_classified.txt -h EU_hier.txt EU_samples.fasta
```

We also want to use RDP's libcompare tool in order to test for significant differences between our two samples. The following command will create the file "libcompare.txt" that we can use for this analysis. Again, note that you will need to enter the correct path to your classifier.

```
java -Xmx1g -jar /path/to/classifier.jar libcompare
   -q1 BF_samples.fasta -q2 EU_samples.fasta -o
   libcompare.txt
```

Statistical analyses in RStudio

In steps 1 and 2 we utilized bash to clone the class repository, preview our files, and run commands for the RDP Tools program. If you'd like to learn more about bash there is a great tutorial **here** provided by Software Carpentry that is geared towards scientists. More helpful links can be found on the links page of this document.

At this point we'll use a different language called R, which was developed for statistical computing and graphics. Many of these things can (and used to) be performed in Excel but R provides a more robust approach, and the program RStudio provides a relatively easy to use graphical interface.

For the rest of this tutorial you'll want to open the file rscript.R in RStudio on your computer. The code is provided for you, with comments explaining each step. During class we'll go over some basics of R and RStudio, and how to run the analysis script provided.

Assignment

The following section will help you frame your understanding of microbial ecology, as well as interpret the results of your analyses and figures. Also included are suggestions for further analysis and/or modifications to the provided R script.

Formulate hypotheses

Answer the following questions, and **include a sentence stating your formal hypothesis** for each.

1. In general, do you think the Europe (EU) communities will be very similar or very different to the Burkina Faso (BF) communities? Why?
2. Do you think the EU communities will differ in the number of types of bacteria compared to the BF communities?
3. Do you think the communities will differ in evenness? Evenness is highest when the taxonomic groups in a sample have the same abundance.
4. Do you think the communities will differ in their function? How?

Results & Interpretation

For the first figure (**taxonomic comparison**):

1. Do the BF and EU communities differ in the number of taxonomic groups at the taxonomic levels you analyzed?
2. Examine your charts. Which community appears more even?
3. Which taxonomic groups are most abundant in the Burkina Faso communities? The Europe communities?
4. How do your results compare to your hypotheses?

For the second figure (**statistical significance**):

1. Which phyla are significantly more abundant in the Burkina Faso communities? The Europe communities?
2. Based on what you know about the physiology of the phyla discussed in the previous question, can you make any inferences about the differences in the function of these communities?
3. How does this relate back to your hypotheses?

Modifications to the R script

There are many ways you can modify the provided R script to gain a better understanding of the R environment, as well as analyze different parts of the data generated with RDPTools.

- Try changing the colors and/or labels of the plots you generated.
  - Hint: This guide will help understand ggplot lables, and this one is helpful for modifying colors.
- Try modifying the data frames generated with dpylyr so you can compare samples at the “class” or “order” level instead of “phylum”. Re-run the script to create new figures that reflect this - but make sure to update the plot labels with the correct taxonomic level!
  - Hint: You shouldn’t need to, but if you want to reshape the data frame here is a handy cheatsheet for dpylyr functions.

Links

Included below are links to resources and tutorials related to the topics covered in this exercise, including some great resources for getting started learning bioinformatics.

Microbial ecology

- The Human Microbiome Project - The National Institute of Health's central repository for human microbiome data and research.
- Berkeley Lab's Microbes to Biomes (M2B) Initiative - Berkeley Lab's microbiome projects focus on both human and soil microbiomes with an emphasis on revealing, decoding, and harnessing microbes.
- UO META Center for Systems Biology - This center at the University of Oregon studies how host-microbe systems work in order to use this knowledge to advance human health.
- Biology and the Built Environment at UO - This center at UO investigates the microbial ecology of indoor environments.
- NASA's Astrobiology Program - Astrobiology is the study of the origin, evolution, distribution, and future of life in the Universe. Microbial ecology is an important part of this, including the study of microbes in extreme environments.

Bioinformatics

General

- UO Applied Bioinformatics and Genomics Master's Program - This is the homepage for the bionformatics master's program at UO and it contains information for prospective students, their annual genomics conference, and more.
- Software Carpentry - This site provides some GREAT lessons for Git, python, R, and many other tools scientists may need to use. They also host live workshop events around the world.
- Rosalind - Focuses mainly on python and includes some beginning tutorials on using python with bioinformatics.
- An Introduction to Applied Bioinformatics - A great, free resource to learning bionformatics from beginner to expert.

R and RStudio

- Swirl: Learn R, in R - This is a great package you can install within RStudio,

serves as a great beginner's tutorial.

- Cookbook for R - IA nice, simply designed site that's great for learning various tricks in R.
- RClub at UO - A very active club right here at the University of Oregon. They have an informative blog, meet often and offer workshops for beginners and intermediate users.