bm211-workshops

Dr Leighton Pritchard and Dr Morgan Feeney

2023-01-01

Table of contents

Preface to the 2023-24 presentation

Welcome to the 2023-24 edition of the BM211 computing workshops.

This is the first presentation of this material in this form, and we would be very grateful to hear feedback by email or through the GitHub repository Issues page.

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

2 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

Part I

Workshop 01

Our goal in this computational workshop is to introduce some concepts in microbial ecology.

3 Microbial Ecology

This section introduces key concepts of how we use molecular techniques to identify and quantify the composition of microbial communities, to study microbial ecology.

> **Note**
>
> This is accompanying material to give additional background and explanation for the workshop. You don't *need* to read this for the workshop itself as we will be covering the material in person, but we hope you find it helpful.

3.1 What is microbial ecology?

Microbial ecology is the scientific study of natural microbial communities. It is a hugely important, but arguably underdeveloped area of biology. Important questions remain to be completely answered in detail for many biological systems:

> The composition of microbial communities
>
> - what microbes are found in a community?
> - what proportion of each microbe species (or other group) is present in a community?
> - what are the physical limits of the community?

> interactions within the community
>
> - is the community *stable* (or is it growing, shrinking, or changing composition)?
> - do community members share or compete for resources, and which resources?
> - do the community members effectively operate as a larger-scale system?

> interactions of a community with other organisms
>
> - does the gut microbial community influence the host animal's nutrition, health, or disease status?
> - does the rhizosphere microbial community influence a plant's ability to extract

nutrition or energy from the environment, and contribute to protection against pathogens?

Interactions between communities

- when two communities come together, what does the resulting community look like?
- are there two or more separate (or interacting communities), or do they somehow combine?

Engineering new microbial communities

- can we predict which microbes can combine into a productive community?
- can we choose these microbes to engineer a community that achieves a particular goal (e.g. plant protection or gut health)
- can we perturb existing communities to make them beneficial to health or achieve some other goal?

This array of questions can be summarised more simply as (Prosser 2020):

1. Who is there?
2. What is meant by "there"?
3. Who is doing what in there?
4. What is the effect of doing … to "there"?

3.2 Answering questions about microbial ecology using molecular techniques

Molecular techniques enabled by modern sequencing approaches, such as amplicon marker sequencing (e.g. 16S, ITS1 methods) - also known as **metabarcoding** - and sequencing the entire genomic DNA of a complex sample - also known as **metagenomics** - have become the dominant approach to understanding microbial community composition (Boughner and Singh 2016). These tools give rapid insights into the complex composition of microbial communities that are unattainable using laborious, time-consuming culture-based techniques.

Metagenomics *vs* Metabarcoding

The two words are quite similar, but it is important to distinguish between

- **metabarcoding**: sequencing a single marker sequence from a population
- **metagenomics**: deep sequencing of the entire genomic DNA material in a sample,

often attempting to assemble representative genomes from the data

In addition to their speed and the scale of data they can produce, these molecular techniques have other advantages:

- The (sequence) data collected can be preserved and shared, alongside metadata describing the experiment, in public databases. This enables reanalysis and integration into larger-scale studies.
- Fastidious organisms, or those that might be outcompeted in a laboratory culture, can be sequenced and identified.
- Organisms present in low numbers can be sequenced and identified.

! Important

These molecular techniques answer the question of *"Who is there?"*, in two ways:

1. They tell us **what kinds of organism are present** (often to *species* or *genus* level)
2. They give us a relative count of **how much of each kind of organism is present**

Taken together, these answers tell us the **diversity of the community**.

i Note

We will introduce measures of community diversity in a later section.

3.2.1 Which organisms are present?

Both **metabarcoding** and **metagenomics** can tell us which microorganisms are present in a biological or environmental sample.

3.2.1.1 Metabarcoding

In **metabarcoding**, we use the polymerase chain reaction (PCR) and **primers targeted specifically to amplify a marker sequence**: a gene fragment or other stretch of DNA that we believe to be present in all organisms of interest. For bacteria, this is usually a fragment of the 16S rRNA gene. For fungi and other eukaryotic mocroorganisms it may be a fragment of the ITS1 (Internally Transcribed Spacer 1) region.

These regions of the genome acquire sequence changes as their organisms evolve over time, such that variants of the marker sequence can be associated with a particular group of microorganisms, usually at the species or genus level. By amplifying and then sequencing the marker sequence from all suitable organisms in the sample, we can **identify the marker sequence variants, and then compare these to databases of known sequences**, to identify which organisms are present.

3.2.1.2 Metagenomics

A similar principle applies with **metagenomics**: we sequence the sample and then compare the resulting sequences to a database containing sequences of known organisms, and use this to identify which of them are present in our sample. Differences between the methods arise mainly because **this approach is untargeted - we do not know in advance what sequences we will recover** - and generates very large amounts of data. Three of the main consequences of this are:

1. We can compare individual sequencing reads to the database of known sequences, but due to the large amount of sequence data (in comparison to metabarcoding) we have to do this using special techniques like ***k*-mer alignment**, rather than direct sequence comparison (Wood and Salzberg 2014).
2. We can assemble near-complete and complete genomes from the data: **Metagenome-Assembled Genomes (MAGs).** (Setubal 2021).
3. We can obtain much more information than simple organism identity, such as the presence or absence of individual genes or gene functions, such as antimicrobial resistance genes (Abreu, Perdigão, and Almeida 2020).

3.2.2 How much of each organism is present?

Whichever method we use to sequence a biological sample, we can obtain a list of the organisms that are present. This tells us something about the composition of the sample, but not everything. There is clearly a difference between one community that is 50% *Escherichia coli* and 50% *Staphylococcus aureus*, and another community that is 99% *S. aureus* and 1% *E. coli*. But how can we distinguish between the two?

In both metabarcoding and metagenomics we can count the number of sequences that correspond to each of the organism types we have identified. For metabarcoding, this might be the count of each amplicon sequence corresponding to a microbial genus. For metagenomics, this might be the number of reads assigned to a microbial species by a tool like Kraken.

⚠ Molecular techniques only *estimate* organism abundance

All counts of organisms, by both methods, are strictly **estimates of the representation of each organism in the sample**. There are many factors that influence the way this estimate may very from the actual amount or proportion of any organism in the sample, including:

- DNA quality (Manzari et al. 2020)
- laboratory technique (sample spillover or other handling problems)
- PCR artefacts (especially for low-biomass samples)
- sequence database integrity
- variable accuracy of associating a sequence with an organism or group (taxonomic misassignment)

With metagenomic analyses that result in assembled MAGs, an additional way of estimating the amount of each organism present is possible: we can count the number of reads that contribute to each MAG assembly. These reads are then assigned the same identity as that of the MAG itself.

We refer to the count of each distinct group of microbes as its **abundance**.

3.3 Describing community composition

The list of organisms identified as present in our sample and the (relative) count of how much of the community is composed of each organism, taken together, describe our measured **microbial community composition**.

4 Diversity Measures

This section introduces the concept of a **diversity measure** in microbial ecology.

4.1 What is diversity?

Diversity measurements for a microbial community are quantitative estimates of how *heterogeneous* or *homogeneous* that community is. *Qualitative* estimates of diversity can be simple. If a community is composed of identical individuals, it is clearly not diverse. However, if the members of a community differ from each other, the community is diverse to some extent.

> **!** Important
>
> Diversity (or the lack of it) is a **property of the community, not a property of individuals** in the community.

If we want to record or compare the level of diversity in a community, we cannot easily rely on qualitative estimates. We need to be more precise than this, and use *quantitative* approaches. We need to know what groups are present in the community and, numerically, to what extent each group is represented.

4.1.1 An example community

To help illustrate this, we went to Mars[1] to obtain a sample (Figure 4.1).

This community contains a range of individuals of different type, distinguishable by the detailed structure of their outer membranes, and also by their internal contents. The diversity of the community can be seen in (Figure 4.2).

[1] OK, we actually went to Sainsbury's Local. The one on Buchanan Street. But the sweets *are* technically from *Mars*.

Figure 4.1: A community sample, obtained from Mars

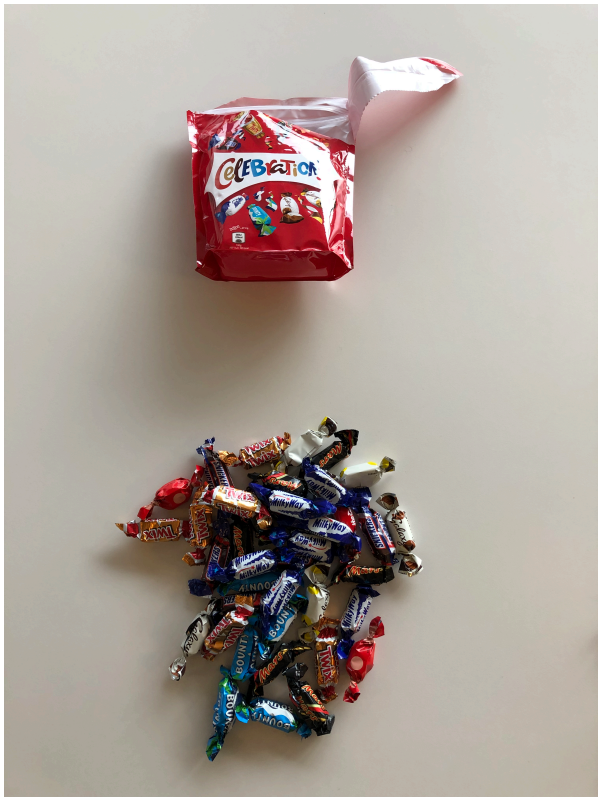

Figure 4.2: A detailed view of the Mars sample, showing community diversity

Exercise 1

By visual inspection of (Figure 4.2), how diverse would you say the sample is?

Answer

There are clearly several different kinds of individual present, so the sample is not homogenous and is *diverse* in some sense. But in another sense the individuals are all individually-wrapped sweets, so the sample is entirely homogeneous.
We clearly need to be more precise about what exactly we mean, when talking about *diversity*.

4.2 Community richness and evenness

We usually define how *diverse* a community is by using a **diversity index**. This is a number that reflects the distribution of different *types* (such as microbial species) of individual that are found in a sample. There are two major concepts to note that contribute to diversity indices.

Richness

The **richness** of a sample is the **number of different types that are represented in the sample**.
For example, *species richness* is a count of the total number of species in the sample. **It does not take into acount species abundance, or relative abundance**.

Evenness

The **evenness** of a sample reflects **how similar is the abundance of each type of thing** in the sample. The more similar the proportion of each type, the more *even* the community.
For example, *species evenness* reflects the relative abundances of distinct species in the sample. If a sample contains three species: A, B, and C, then:

- a sample with 33% A, 33% B, and 33% C has **high species evenness**
- a sample with 5% A, 90% B, and 5% C has **low species evenness**

4.3 Richness and evenness in the Mars sample

We can organise the individuals in the Mars sample, as shown in Figure 4.3. This gives us enough information to calculate species richness and evenness for this community.

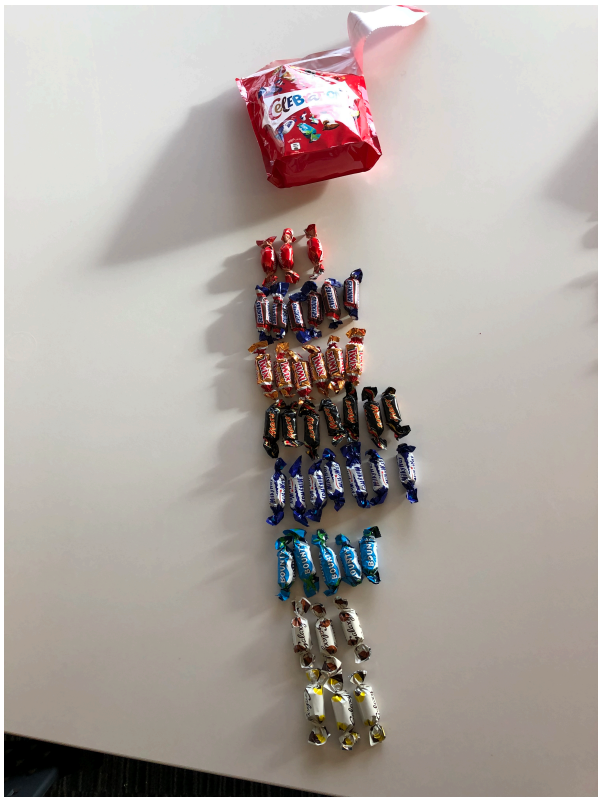

Figure 4.3: Species abundance in the Mars sample community

4.3.1 Species richness

We can easily count the number of different species, as determined by morphology and membrane features, in (Figure 4.3). There are eight (8).

Taken in isolation, this number doesn't tell us very much. We would need other information - maybe from other samples - to know whether eight species is a large, small, or intermediate number of species. Nor does this number tell us whether all species are represented to the same extent (even if, from looking at Figure 4.3 we can tell they are not).

4.3.2 Species diversity: Shannon index

To understand diversity, we need to consider both richness and evenness. We know that if each species is present in about the same amount there is *high evenness*, and if the population is dominated by a single species there is *low evenness*, but by itself that does not give us a number that we can use to compare communities. But there are formulae that allow us to do this, and we can calculate the value using `R`.

4.3.2.1 Calculating Shannon Index

Let's start in `R` by defining our dataset for the Mars community. In the code below we do this by creating two vectors: one of species names, and one of counts for those species in our sample. We combine these vectors into a single dataframe, for convenience.

```
# Define a vector of species names
species <- c("Malteaser sp.", "Snickers sp.", "Twix sp.", "Mars sp.",
             "Milky way", "Bounty sp.", "Galaxy choc", "Galaxy caramel")

# Define a vector of corresponding species counts
count <- c(3, 6, 6, 7, 7, 5, 3, 3)

# Bring these together in a dataframe
community.df <- data.frame(species, count)
```

We can inspect the dataframe, to make sure we have the correct data:

```
community.df
```

```
         species count
1  Malteaser sp.     3
2   Snickers sp.     6
3       Twix sp.     6
4       Mars sp.     7
5      Milky way     7
6     Bounty sp.     5
7    Galaxy choc     3
8 Galaxy caramel     3
```

The first step in calculating Shannon Index is to calculate the **relative abundance** of each species. To do this, we calculate the percentage of the entire community that is made up from

each species, as below. We divide the count for each species by the sum of counts for all species (i.e. the total number of individuals, in this case).

```
# Calculate relative abundances
community.df$rel_abd <- community.df$count / sum(community.df$count)

# Show the data frame
community.df
```

```
         species count rel_abd
1  Malteaser sp.     3   0.075
2   Snickers sp.     6   0.150
3       Twix sp.     6   0.150
4       Mars sp.     7   0.175
5      Milky way     7   0.175
6     Bounty sp.     5   0.125
7    Galaxy choc     3   0.075
8 Galaxy caramel     3   0.075
```

To turn this data into the Shannon Index, we need to carry out two more steps: calculate the *natural log* of the relative abundance of each species, then multiply this by the corresponding relative abundance, as in the R code below:

```
# Calculate the natural log of relative abundance
community.df$ln_rel_abd <- log(community.df$rel_abd)

# Multiply the relative abundance by its natural log
community.df$mult <- community.df$rel_abd * log(community.df$rel_abd)

# Show the result
community.df
```

```
         species count rel_abd ln_rel_abd       mult
1  Malteaser sp.     3   0.075  -2.590267 -0.1942700
2   Snickers sp.     6   0.150  -1.897120 -0.2845680
3       Twix sp.     6   0.150  -1.897120 -0.2845680
4       Mars sp.     7   0.175  -1.742969 -0.3050196
5      Milky way     7   0.175  -1.742969 -0.3050196
6     Bounty sp.     5   0.125  -2.079442 -0.2599302
7    Galaxy choc     3   0.075  -2.590267 -0.1942700
8 Galaxy caramel     3   0.075  -2.590267 -0.1942700
```

The Shannon Index is then the sum of this final column of values (multiplied by `-1` to make it a positive value).

```
shannon_index = -sum(community.df$mult)
shannon_index
```

```
[1] 2.021916
```

Part II

Workshop 04

Our goal in this workshop is to introduce you to handling and analysing CRISPR-Cas sequence data.

PILER-CR

This chapter will introduce the software tool `PILER-CR`, which is designed to predict the location and content of CRISPR repeats in microbial genomes.

Part III

Workshop 05

Our goal in this workshop is to introduce you to producing and interpreting phylogenetic trees.

Protein Trees

This chapter will guide you through the process of producing a phylogenetic tree from a protein sequence alignment.

References

Abreu, Vinicius A C de, José Perdigão, and Sintia Almeida. 2020. "Metagenomic Approaches to Analyze Antimicrobial Resistance: An Overview." *Front. Genet.* 11: 575592. https://doi.org/10.3389%2Ffgene.2020.575592.

Boughner, Lisa A, and Pallavi Singh. 2016. "Microbial Ecology: Where Are We Now?" *Postdoc J.* 4 (11): 3–17. https://doi.org/10.14304%2FSURYA.JPR.V4N11.2.

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.

Manzari, Caterina, Annarita Oranger, Bruno Fosso, Elisabetta Piancone, Graziano Pesole, and Anna Maria D'Erchia. 2020. "Accurate Quantification of Bacterial Abundance in Metagenomic DNAs Accounting for Variable DNA Integrity Levels." *Microb. Genom.* 6 (10). https://doi.org/10.1099%2Fmgen.0.000417.

Prosser, James I. 2020. "Putting Science Back into Microbial Ecology: A Question of Approach." *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375 (1798): 20190240. https://doi.org/10.1098/rstb.2019.0240.

Setubal, João C. 2021. "Metagenome-Assembled Genomes: Concepts, Analogies, and Challenges." *Biophys. Rev.* 13 (6): 905–9. https://doi.org/10.1007/s12551-021-00865-y.

Wood, Derrick E, and Steven L Salzberg. 2014. "Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments." *Genome Biol.* 15 (3): R46. https://doi.org/10.1186/gb-2014-15-3-r46.