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**.

Diversity Measures

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

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.