An Introduction to Metagenomics and Microbial Ecology

Ilia Korvigo

This document is a supplement to an introductory course on microbial ecology and metagenomics taught at the Saint Petersburg State University and the ITMO University. It comprises a concise summary of lecture materials, practical guides and home-work assignments.

Contents

1 Introduction

1.1 Microbial ecology and metagenomics

Microbial ecology is the study of microbiomes, that is consortia of arbitrarily diverse microbial populations interacting with each other and the environment they inhabit. Metagenomics is an emergent field within computational biology comprising a loose grouping of experimental, computational and statistical methods dedicated to generating and analysing highly heterogeneous genetic information (i.e. mixtures of genomes originating from distinct organisms). While we should not confuse microbial ecology with metagenomics (as is often done in contemporary parlance), metagenomics has become the de facto standard way to study microbial communities, inasmuch as genetic information (be it DNA or RNA) is the only type of information we can extract from said communities efficiently enough. In other words, modern microbiome research is deeply grounded in metagenomics, though the latter has other applications. Metagenomic data come in two flavours: shotgun (aka "complete") metagenomes and amplicon libraries. As their names suggest, the difference boils down to the amount (or type) of genetic information we sample from an environment. Although shotgun metagenomics is becoming increasingly accessible and viable (though not enough to become truly practical in most cases), amplicon libraries remain the most common type of metagenomic data (whilst not being metagenomes *sensu stricto*) due to an unparalleled combination of cost efficiency, community coverage and analytical convenience, which is why this course only covers amplicon metagenomics.

1.2 Amplicon sequencing

Unlike shotgun metagenomes, amplicon libraries comprise different instances of a single target DNA sequence. Any sequence is a valid target as long as we can design a pair of universal primers compatible with all its instances within a selected phylogenetic group (be it large or small). Targeting a specific piece of DNA instead of entire genomes greatly reduces sequencing costs, downstream computational requirements and improves diversity coverage (especially in low-entropy communities) at the expense of direct functional information (unless the target itself determines the function of interest). Whether the target is a sequence pertaining to some functional activity (e.g. an enzyme or a receptor) or a universal marker-gene with no immediate functional implications (e.g. 16S rRNA), it contains phylogenetic information about the underlying phylogeny of captured community com-

ponents. Students are highly encouraged to read [1] and [2] to get a better grasp of phylogenetic information as a concept, though for now we can make do with a simple rule of thumb: a target sequence should be as long as possible and neither too conserved, nor too variable.

1.3 The 16S rRNA gene

The 16S rRNA gene is by far the most common target for amplicon metagenomics. It is not only truly omnipresent, but also conserved enough to enable universal primers compatible with all known prokaryotic sequences (as well as some mitochondrial and plastid sequences and even 18S eukaryotic genes [3]). The gene consists of short conserved patches spread between 9 longer hyper-variable regions. The most prominent of these are hyper-variable regions V3 ($\sim$ 180 bp) and V4 ($\sim$ 290 bp) Most of the older microbiome studies targeted either of the two, while newer studies are increasingly targeting both regions (i.e. the V3-V4 fragment) thanks to widespread adoption and improved quality of 250 bp and 300 bp Illumina paired-end sequencing.

2 Preprocessing

2.1 Problem statement

An amplicon sequencing library is a noisy collection of extracted, amplified and sequenced instances of a target DNA/RNA sequence. In most practical cases we do not know exactly what instances to expect — we have to infer them from the data, while accounting for artefacts, such as distorted and chimeric sequences, to avoid diversity inflation. Then, we need to transform the library into a count vector $\mathbf{x} = (x_i, \cdots, x_n)$, where x_i is the number of observations (reads) corresponding to instance i. Bear in mind, that all steps in an experimental metagenomic workflow, from DNA extraction and purification to sequencing itself, inevitably distort the underlying community structure in one way or another, and there is little we can do about it.

2.2 Clustering-based preprocessing

Up until recently, the absolute majority of published microbiome studies have used a data preprocessing step known as OTU-picking. In general, an OTU (operational taxonomic unit) is a group of target sequences satisfying a similarity criterion. There are three main types of OTU-picking strategies: de novo, closed-reference and open-reference (the last one is a combination of the other two). De novo OTU-picking algorithms create groups from scratch and select (or compute) one representative sequence (a centroid) per group. In closed-reference OTU-picking we start with an existing set of representative sequences (references) and group reads from an amplicon library with any one of the references according to a similarity criterion without creating any new representatives. In practice, most OTU-picking algorithms involve some sort of alignment-based sequence clustering with a fixed similarity threshold (e.g. 97%). The process is usually preceded by primer excision (so as not to bias similarity scores), quality-score-based trimming and read-pair merging and is followed by chimera removal and frequency-based filtering.

A fixed similarity threshold is supposed to absorb experimental variations and group all inexact readings of a single underlying target instance. However, it also incurs a risk of merging actual distinct targets, thereby deflating diversity estimates and erasing potentially valuable statistical information. In addition to that, accurate (de novo) clustering (based on complete-linkage algorithms) is not computationally feasible, while heuristic alternatives (including advances statistical methods with adaptive similarity thresholds, e.g. [4]) are extremely unreliable [5, 6, 7]. Finally, the inability to jointly analyse

datasets produced by independent OTU-picking runs (with the exception of closed-reference OTU picking) adds to the list of reasons to avoid OTU-picking altogether.

2.3 Denoising-based preprocessing

OTU-picking has largely fallen out of favour with progressive microbiome researchers (the rest are catching up rapidly). Denoising reads (that is applying error-correction) is a far better and consistent way to handle experimental variation in sequencing data, because it does not involve any similarity-based clustering. There are many different denoising algorithms, which is why students are advised to read a comparative review [8] of the three most popular ones, namely DADA2, Deblur and UNOISE3. In any case result is a set of all unique error-corrected target instances observed in an amplicon sequencing library, that is amplicon sequencing variants (ASVs) also known as sub-OTUs.

2.4 Chimeric sequences

A chimera is a synthetic combination of two or more distinct DNA templates. They appear when an aborted PCR product of one real DNA template anneals to a different template during multi-template PCR. Most amplicon sequencing libraries contain chimeras, though the exact fraction of chimeric sequences varies widely and depends on multiple factors, such as initial template concentrations and the total number of PCR cycles. It is not entirely unusual to have chimeras amount to over 50% of a sample. Consequently, all decent amplicon data preprocessing workflows include chimera detection and removal (typically as the final step). Although perfect chimera removal is does not seem feasible even when we possess exact sequences of all initial DNA templates [], the choice of upstream preprocessing steps does influence the accuracy a lot. In particular, OTU-picking hinders chimera-detection by masking non-experimental variation in closely related initial (i.e. real) template sequences. Since denoising strategies tend to reconstruct most of the real initial templates, they greatly improve and simplify chimera detection.

2.5 Copy-number variations

A single genome can encode multiple instances of a target sequence: identical (or sufficiently similar) copies distort component abundances, while distinct copies appear as separate entities (as though sequences originating from different microbial populations) and distort community diversity. Considering

that, the number of ribosomal operons varies a lot (in prokaryotes as well as eukaryotes) even on a small phylogenetic scale, copy-number-variations are another significant data-obfuscating factor to be aware of, even though there is not much we can do about it now (if ever).

2.6 Assignment

Directory `${HOME}/course/assignment-one/noprimer` contains a single paired-end 16S amplicon sequencing library after primer excision. Your must:

- Write a bash-script for OTU-picking-based preprocessing with de novo clustering:
  1. Quality-score-based trimming (`trimmomatic`)
  2. Merge pairs (`fastq-join`)
  3. Convert merged fastq reads into a QIIME 2 artefact (`qiime tools import`)
  4. Dereplicate sequences (`qiime vsearch dereplicate-sequence`)
  5. Cluster sequences (`qiime vsearch cluster-features-de-novo`)
  6. Run de novo chimera detection (`qiime vsearch uchime-denovo`)
  7. Exclude chimeras (`qiime feature-table filter-features`)
  8. Export a feature table (OTU counts) and OTU sequences
- Write an R-markdown notebook for denoising-based preprocessing using DADA 2

All software packages required to complete the task are available in the `student` conda virtual environment.

3 Taxonomic annotation

3.1 Problem statement

From a practical point of view, taxonomic annotation is an implicit way to augment amplicon sequencing data with functional information based on microbiological knowledge about certain taxonomic groups. However, taxonomy is a largely synthetic construct that can easily defy "natural" phylogenetic relationships between microbial populations. This is especially true about prokaryotic taxonomy, because, on the one hand, standard (i.e. eukaryotic) definitions of species (based on reproductive isolation of populations) are incompatible with prokaryotic modes of reproduction, and, on the other hand, traditional prokaryotic systematics is more concerned with phenotypic/functional traits than actual phylogeny. For example, *Shigella* and *Escherichia* are treated as two separate genera, though phylogenetically *Shigella* species fall well within the natural intra-species variation of *E. coli* [9, 10]. This phylogenetic inconsistency makes accurate taxonomy prediction (and, by extension, functional inference) almost impossible, especially given a rather minuscule amount of genetic information we have in short-read amplicon sequencing datasets.

3.2 Reference datasets

Taxonomic annotation requires a set of reference sequences with known taxonomic identities. There are three large general-purpose (that is domain-agnostic) 16S rRNA datasets: RDP, Silva and GreenGenes; only the first two of them are updated on a regular basis. All three databases are error-ridden and thus require careful filtering to improve annotation accuracy [11]. Sequences with partial, conflicting or completely inaccurate taxonomic annotations constitute up to a quarter of all records. Manually curated domain-specific databases, such as HITdb, tend to be significantly more accurate, but they are rarely available.

3.3 Alignment-based annotation

Given a database of reference sequences with known taxonomic identities, sequence alignment becomes the most straightforward approach to taxonomic annotation. This process is essentially similar to closed-reference OTU-picking with a couple of noticeable distinctions. First of all, alignment-based taxonomic annotation is usually applied to preprocessed data (i.e. chimera-checked and filtered ASVs or OTUs), though it is not entirely uncommon to

keep taxonomic annotations produced as a side-product of a closed-reference OTU-picking run. Second of all, accurate alignments are computationally feasible, because we only need to align ASVs or OTU centroids. As with OTU-picking, the 97% similarity threshold was historically considered a good criterion for a perfect reference match. However, modern studies show that only 100% similarity provides any sort of annotation accuracy for partial 16S sequences (e.g. the V4 region) [12, 13]. The bigger problem is the inability to control annotation uncertainty.

3.4 Machine-learning-based annotation

Machine-learning-based taxonomic classifiers typically provide at least some way of controlling prediction uncertainty. The most popular taxonomic classifiers are the RDP classifier [14] and IdTaxa [15]. The first one is a naïve Bayesian classifier, and the second one implements a custom model inspired by decision trees. Both classifiers treat sequences as bags-of-words, that is unordered k-mer spectra. Since k-mer spectra of partial 16S sequences diverge from complete-sequence spectra, to avoid unreliable uncertainty estimates we must always use classifiers trained on the target regions we are dealing with.

3.5 Phylogenetic placement

Given a phylogenetic tree of annotated referency sequences, phylogenetic placement algorithms determine the most probable location of a query sequence under a maximum likelihood or a Bayesian phylogenetic model. Although placement is quite an old method, it has become far more accessible (i.e. user-friendly) with the advent of SEPP [16], a single-command plugin for the QIIME 2 amplicon data analysis framework that comes bundled with a precomputed GreenGenes phylogenetic tree. Phylogenetic placement is in many ways superior (at least theoretically) to both alignment-based and machine-learning-based methods. In practice, here the accuracy depends not only on the quality of reference annotations (as is mostly the case with the other two methods), but also on the quality of reference phylogeny; and large-scale phylogenetic trees are never really accurate [11] and extremely computationally expensive to infer.

3.6 Assignment

Directory `${HOME}/course/assignment-two` contains a symlink `silva.fna` to a database of reference 16S rRNA sequences. Sequence descriptions con-

tain taxonomic annotations (normal prokaryotic sequences have a 7-level taxonomic description). Another file, `primers.fna`, contains a pair of universal 16S rRNA primers. You must:

- Write a Python Jupyter or an R-markdown notebook to preprocess the reference dataset:
    1. Convert sequences from RNA to DNA;
    2. Exclude non-prokaryotic sequences;
    3. Extract the target region (i.e. the region bound by the primers);
    4. Exclude sequences that failed target extraction;
    5. Investigate all taxonomic ranks and devise a set of rules to remove sequences with incomplete taxonomic annotations;
    6. Apply these rules;
    7. Group taxonomies by extracted target sequences and summarise the number of groups with taxonomy conflicts at each taxonomic level;
    8. Export each unique Taxonomy-Target pair into a FASTA file consistent with the sequence-labelling format of the original SILVA file;
- Write an R-markdown notebook to train a genus-level IdTaxa taxonomic classifier;
    1. Group all target sequences by Phylum-Class-Order-Family-Genus labels;
    2. Down-sample overrepresented genus groups to allow no more than 10 sequences per genus;
    3. Train a classifier for 20 iterations;
    4. Save the classifier;
- Write an R-markdown notebook to train a species-level IdTaxa taxonomic classifier;
    1. Group all target sequences by Phylum-Class-Order-Family-Genus-Species labels;
    2. Down-sample overrepresented species groups to allow no more than 10 sequences per species;
    3. Train a classifier for 20 iterations;

4. Save the classifier;

- Apply both classifiers to predict taxonomy for OTUs and ASVs produced during the previous homework assignment (confidence level: 80%);

References

[1] M. Bordewich *et al.*, "On the information content of discrete phylogenetic characters," *Journal of Mathematical Biology*, vol. 77, pp. 527–544, 2018.

[2] P. O. Lewis *et al.*, "Estimating Bayesian Phylogenetic Information Content," *Systematic Biology*, vol. 65, pp. 1009–1023, 2016.

[3] Y. Wang *et al.*, "Optimal Eukaryotic 18S and Universal 16S/18S Ribosomal RNA Primers and Their Application in a Study of Symbiosis," *PLoS ONE*, vol. 9, p. e90053, 2014.

[4] S. W. Olesen *et al.*, "dbOTU3: A new implementation of distribution-based OTU calling," *PLOS ONE*, vol. 12, p. e0176335, 2017.

[5] W. Chen *et al.*, "A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs," *PLoS ONE*, vol. 8, p. e70837, 2013.

[6] M. A. Jackson *et al.*, "A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units," *PeerJ*, vol. 4, p. e2341, 2016.

[7] R. C. Edgar, "Accuracy of microbial community diversity estimated by closed- and open-reference OTUs," *PeerJ*, vol. 5, p. e3889, 2017.

[8] J. T. Nearing *et al.*, "Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches," *PeerJ*, vol. 6, p. e5364, 2018.

[9] J. R. Johnson, "Shigella and Escherichia coli at the crossroads: machiavellian masqueraders or taxonomic treachery?," *Journal of Medical Microbiology*, vol. 49, pp. 583–585, 2000.

[10] E. A. Pettengill *et al.*, "Phylogenetic Analyses of Shigella and Enteroinvasive Escherichia coli for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation," *Frontiers in Microbiology*, vol. 6, p. 1573, 2016.

[11] R. Edgar, "Taxonomy annotation and guide tree errors in 16S rRNA databases," *PeerJ*, vol. 6, p. e5030, 2018.

[12] R. C. Edgar, "Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences," *PeerJ*, vol. 6, p. e4652, 2018.

[13] R. C. Edgar, "Updating the 97% identity threshold for 16S ribosomal RNA OTUs," *Bioinformatics*, vol. 34, pp. 2371–2375, 2018.

[14] Q. Wang *et al.*, "Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy," *Applied and Environmental Microbiology*, vol. 73, pp. 5261–5267, 2007.

[15] A. Murali *et al.*, "IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences," *Microbiome*, vol. 6, p. 140, 2018.

[16] S. Janssen *et al.*, "Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information," *mSystems*, vol. 3, pp. e00021–18, 2018.