Optimal metabarcoding - Statistical analysis

Emma Granqvist & Fredrik Ronquist

June 12, 2022

1 Statistical analysis

In this section, our goal is to develop a probabilistic model that allows us to learn parameters of interest from the observed read counts in a metabarcoding analysis of a community sample. First, we will use the data from the controlled experiments reported here in analyzing critical model parameters, such as the variation in DNA yield among specimens and between extraction protocols, and the variation in PCR bias among species. We will then explore how accurately we can predict the number of specimens or biomass of different species in an unknown sample, given that we have data on spike-ins, that is, foreign species that have been added to the sample in known quantities for calibration purposes.

1.1 Notation

We first summarize our notation (see also Table in Appendix). Consider a set of community samples indexed by $m = \{1, 2, ..., M\}$. Each community contains a subset of a global set of T species (or OTUs); index each species with $t = \{1, 2, ..., T\}$. Finally, index each individual specimen of a species t in a community sample m with $i = \{1, 2, ..., n_{tm}\}$, where n_{tm} is the total number of specimens of that species in the sample. Each community sample is subjected to amplicon-based metabarcoding in one or more samples that we index with $j = \{1, 2, ..., J\}$, where J is the total number of samples for all communities. The observational data consist of read counts for each species t in each sample j of a community m; denote these r_{mtj}.

1.2 Basic model

Assume that the amount of DNA we extract from an insect specimen i belonging to species t in sample j of a community m is d_{mtij} per volume we use for the PCR reaction. We call this the *DNA yield*. The DNA yield may be affected by the sample treatment (e.g. lysis time, homogenization and buffer choice), the general properties of the species, and the properties of the particular specimen. It is convenient for our purposes to model the DNA yield using a gamma distribution, because it allows us to easily accommodate these effects.

In the simplest case, assume that the DNA yield is not affected by the species, the community sample or the extraction protocol. Then we have

$$d_{tmij} \sim \text{Gamma}(k, \theta), \tag{1}$$

where k is the shape parameter and θ is the scale parameter. The parameter d_{tmij} has a natural interpretation: it is the expected number of template DNA copies coming from an individual insect specimen and available for the PCR reaction.

If there are n_{tm} specimens of the species t in the community m, then the total number of template copies will distributed as the sum of n independent draws from the same gamma distribution. This is another gamma distribution, with the same scale parameter but with the shape parameter $n_{tm}k$. That is,

$$d_{tmj} \sim \text{Gamma}(n_{tm}k, \theta). \tag{2}$$

We model the PCR reaction, the sequencing and the bioinformatic processing as a multiplication of the DNA yield by a factor c_j, which is specific for the sample j but multiplies all template DNA in the same way. In other words, $r_{tmj} = c_j d_{tmj}$. We call c_j the *PCR factor*. Using the basic properties of the gamma distribution, we know that such a multiplication results in a new gamma distribution with the same shape parameter but with a new scale parameter $c_j\theta$. That is,

$$r_{tmj} \sim \text{Gamma}(nk, c_j\theta). \tag{3}$$

Now, given suitable prior probability distributions on k, θ and c_j, these parameters can be inferred (learned) from observed read counts using probabilistic (Bayesian) inference. This is our basic framework.

1.3 Priors

The chosen priors for the parameters are listed below. We assume a Gamma prior on k. For θ and c we use a log-normal prior, and since θ should be a relatively small number its prior is wide on small values. To estimate reasonable prior settings for c, it is assigned a hyper-prior; we use a conjugate normal-inverse-gamma prior for the parameters of the log-normal distribution for c: μ_c and σ_c. The mean of the hyperprior is set to 6.5, which is close to the mean of the log of the read counts. This is consistent with a mean for the DNA yield of 0.0.

$$k \sim Gamma(shape : 1, scale : 10) \tag{4}$$

$$theta \sim \log\mathcal{N}(0.0, 2.0) \tag{5}$$

$$c \sim \log\mathcal{N}(\mu_c, \sigma_c) \tag{6}$$

1.4 Extensions

Our basic framework can easily be extended. For instance, we can examine whether there are species-specific differences in DNA yield, which will result in species-specific variation in the shape parameter k. Such variation could be due to different body mass of different species (if DNA yield is proportional to body mass), or that different species leak DNA into a mild lysis buffer at different rates.

If there are species-specific differences in the effectiveness of the PCR reaction, so-called PCR bias, we assume that this will manifest itself in the form of a species-specific multiplicative effect on the PCR factor. This will modify the scale parameter of the gamma distribution that describes the number of read counts.

Assuming that one will extract different amounts of DNA from each species of insect, which would mean that k would vary between species:

$$r_{tij} \sim Gamma(k_t, theta * c_j), \tag{7}$$

Assuming that the PCR-factor will vary between species, which would result in *theta* varying between species so that:

$$r_{tij} \sim Gamma(k, theta_t * c_j). \tag{8}$$

1.5 Estimating the composition of samples from spike-ins

We also analyze the performance of our modeling framework in estimating the composition of unknown community samples. We first train the model on data from a set of spike-ins, that is, foreign species that have been added in known quantities to the analyzed samples. Specifically, we estimate the parameters of the DNA yield distribution and the analysis-specific PCR factors c_j. We then use the fact that the read count of an unknown species in a particular sample is drawn from a mixture of gamma distributions with different shape parameters, and we predict the number of specimens (or biomass) of the species by identifying the component of the mixture that is most likely to have generated the observed number of reads.

2 Appendix

2.1 Key symbols

- i = index of insect specimen
- m = index of community
- j = index of sample
- t = index of species
- d_{tij} = amount of DNA extracted per volume used for PCR reaction (DNA yield) for specimen i of species t in sample j
- r_{tj} = the number of reads (read count) for species t in sample j
- c_j = PCR amplification factor for sample j (includes any dilution factors involved)
- k and θ = shape and scale parameters of the gamma distribution of DNA yield
- μ and σ = mean and standard deviation (on the log scale) of the lognormal distribution of PCR amplification factors