Model description of metabarcoding analysis of CALCOFI samples

Motivating models.

We use metabarcoding approaches outlined above to produce amplicons from ethanol-preserved icthyoplankton samples. Briefly, we generated amplicons using the MiFish **12S** Universal Teleost primer set on DNA extractions derived using the Qiagen DNeasy Blood and Tissue kit filtered ethanol-preservative. Each amplicon library was sequenced separately on an Illumina NextSeq.

We estimate that, for any species i, the number of sequenced amplicons is proportional to the fraction of DNA from that species in the PCR template. The amplicons produced during a PCR reaction are governed by the efficiency parameter a_i, which is characteristic of the interaction between the particular primer set and each species being amplified. Thus, for any species i, the number of amplicons should be directly related to the efficiency of amplification and the starting concentration of DNA template such that

$$A_i = c_i(1 + \alpha_i)^{N_{PCR}} \tag{1}$$

where A_i is amplicon abundance, c_i is the true number of DNA copies in the reaction attributable to species i, α_i is the amplification efficiency (bounded on $(0, 1)$), and N_{PCR} is a known constant (an integer giving the number of PCR cycles used in the reaction). If we could perfectly observe amplicons, the above equation would be all that we needed to relate the Amplicons we observe and the biological value of interest, c_i, the true number of template DNA copies. Unfortunately PCR and sequencing technology does not allow for such direct observation. Due to N_{PCR} being a large number and α_i typically being not close to 0, the number of amplicons expected for any species with $c_i > 0$ is very, very large (e.g. with $c_i = 2$, $\alpha_i = 0.75$, and $N_{PCR} = 36$, $A_i = 1.12 \times 10^9$) and there are typically many species being amplified simultaneously, producing 10^{10} or more DNA copies in a single reaction. The actual number driven primarliy by the α values among species and N_{PCR}. This model assumes that PCR amplification has not approached saturation and therefore the PCR is still amplifying exponentially.

DNA sequencing machines do not read all of the copies from such a reaction; they read only a small fraction of the reads (on the order of 10^6 to 10^7 reads). This subsampling changes what in eq. 1 appears to be a single-species process – each species being amplified independently – into a multi-species process; the number of amplicons observed for species i will depend upon both the amplicons produced for species $i = 1$ and the amplicons produced for species $i = 2, 3, ..., I$ in the same reaction. Observations of amplicons are thus compositional data and need to be analyzed as such.

Models for compositional data

We want to retain the data-generating structure from eq.1 as much as possible, so we develop a model for a single sample with many species. As above, let i index species with $i = 1, 2, ..., I$ and then we can write a deterministic equation for the number of amplicons observed in log-space as

$$\log(A_i) = \log(c_i) + N_{PCR} \log(1 + \alpha_i) + \log(\eta) \tag{2}$$

where the only new term is η which represents the proportion of reads observed from a given sampling run. Note that in this formulation η is a single value shared across all species and serves to scale the number of amplicons observed. Additionally we can rewrite the number of DNA copies in terms of proportional number of counts, $\log(\beta_i) = \log(c_i) - \log(\sum_i c_i)$. Note that the second term in this equation is a sum of the counts across all species, and so is a single shared value for all species. As such it can be absorbed into the value η that scales the overall abundance,

$$\log(A_i) = \log(\beta_i) + N_{PCR} \log(1 + \alpha_i) + \log(\eta) \tag{3}$$

This equation is appealing because it provides a process-oriented description of the biology of interest (the βs), a term for how PCR modifies our observation of the true abundance ($N_{PCR}\log(1+\alpha_i)$), and a term for how subsampling of DNA reads in the sequencing machine will adjust the number of amplicons observed $\log(\eta)$. This third term also links all of the single-species components to produce a multi-species model. It is important to note that while both eq. 2 and 3 use the term η, the interpretation and plausible range of this parameter are distinct in the two equations. In eq. 2, $0 < \eta \leq 1$, while in eq. 3 η is not constrained to be less than 1 ($\eta > 0$).

In practice, PCR and subsampling are not deterministic but random processes. Furthermore,we are rarely interested in results from a single sample but rather multiple samples collected across sites j and times t. We let λ_{ijtk} be the expected number of amplicons observed, with k indexing unique PCR reactions to account for the fact that there may be multiple individual PCR reactions for a single collected sample,

$$\log(\lambda_{ijtk}) = \log(\beta_{ijt}) + N_{PCR}\log(1+\alpha_i) + \log(\eta_{jtk}) \tag{4}$$

Importantly, α_i is assumed to be constant for each species among all sites, times, and PCR reactions. We incorporate stochasticity by allowing to the number of observed amplicons to vary from the deterministic mean by specifying the observed values as following a negative binomial distribution,

$$Y_{ijk} \sim NegativeBinomial(\lambda_{ijtk}, \phi) \tag{5}$$
$$\phi = \exp[\tau_0 + \tau_1 \log(\lambda_{ijtk})] \tag{6}$$

where the expected value and variance of Y_{ijk} are $\boldsymbol{E}[Y_{ijk}] = \lambda_{ijtk}$, and $Var[Y_{ijk}] = \lambda_{ijtk} + \frac{\lambda_{ijtk}^2}{\phi}$, respectively. Note that we allow for the scale parameter ϕ to vary with the predicted mean, this allows for the amount of dispersion in the negative bionomial to shift with changing predictions. For our datasets, this allows dispersion to be large when λ is small and decrease as λ increases.

By itself, this model has substantial identifiability problems; in the absence of additional information, it is not possible to estimate the β and α parameters. In the next section we discuss how to integrate data from amplicon sequencing as well as other data sources to make the parameters identifiable.

Application to the CalCOFI Dataset

At each CalCOFI station, an oblique bongo-net tow is conducted from 210 m to the surface with the starboard side preserved in buffered formaldehyde and the port side preserved in buffered ethanol-preservative (detailed above). Manual counts of ichthyoplankton were quantified using microscopy to identify species abundance from formaldehyde-preserved samples. Metabarcoding was conducted on the ethanol preserved side; consequently, we expect the contents of the paired samples to differ slightly as a function of sampling stochasticity. Counts of larvae/juveniles were done once for each jar.

The manual counts provide extra information that enable us to estimate the counfounded parameters from the metabarcoding. Specifically, for each sampled station, we have two sets of observed data: 1) counts of larval/juvenile fishes for each taxon from the formaldehyde jars (Z_{ijt}; indexes as above) and 2) counts of amplicons for each taxon from ethanol jars (Y_{ijtk}). These observed data arise from a common (but unobserved) biomass for each species at each station-year combination (γ_{ijt}; a latent (unobserved) variable).

We assume that the data are counts for each species in each sample, Z_{ijt}, derived from the true density of each species γ_{ijt}, the fraction of total specimens counted in each vial, P_{jt}, and the volume of water filtered for that station relative to a standard volume, V_{jt}; $V_{jt} \approx 1$ for most samples, $V_{jt} < 1$ indicates a smaller volume of water was sampled.

$$Z_{ijt} \sim Poisson(\theta_{ijt}) \tag{7}$$
$$\log(\theta_{ijt}) = \log(\gamma_{ijt}) + log(P_{jt}) + \log(V_{jt}) \tag{8}$$

We consider β_{ijt} to be the true proportion of biomass at a given station-year for each taxon i, $\beta_i = \frac{\gamma_{ijt}}{\sum_i \gamma_{ijt}}$.

Joint Model for Counts and Amplicons

To combine our information from the manual counts and metabarcoding, we need to recognize that our observations (Y_{ijtk} and Z_{ijt}) are linked together by a common variable (γ_{ijt}) and thus we can model them jointly. We represent the amplification process using equation 5 and 6 above (amplicons were sequenced in triplicate reactions for each jar). The manual count are modeled as in equations 7 and 8.

Our model assumes the fraction of template DNA in each sample is proportional to the counts of individual species' larvae in each paired jar. Moreover, we assume that in each sample there is template DNA from species that are uncounted, unidentifiable, or otherwise unobserved. In practice, this DNA might derive from stochastic sampling between each side of the bongo net, excreted waste, stray tissue, cells, or microscopic genetic material extracted along with the visible larvae.

Dealing with the fact that not all methods see the same species

The above is sufficient if all of the species identified by morphological counts are identical to the species identified by the genetic methods. But this is often not the case; some larvae are not separable to species based on morphology and some species are not separable to species based on a single genetic primer. Furthermore, some species do not amplify at all in the PCR ($\alpha_i \approx 0$). To accommodate non-overlaping sets of species among sampling methods we introduce a new variable, γ_{Mijt}, which specifies the true (M is for "main") density of species i at site j and time t. We assume that there is a mapping between this main density and the density observed by each sampling method. Specifically, we assume the species in the main list maps uniquely on to no more than one taxonomic group in each observation method, but multiple main species can map onto a single group for each observation method. For example, if the observation of larval counts identified as a specimen as *Sebastes* sp., we assume this may map onto one or more specific taxa (e.g., *Sebastes paucispinis*) in the main list, but conversely, *Sebastes paucispinis* on the main list may not map to more than one entity identified by each observation method.

We can construct a mapping matrix, $\mathbf{M}_{MS}$, that transforms the species in the main list, $\boldsymbol{\gamma}_M$ (a vector of length I_M, the number of true species in the sample) into the species grouping observed by sampling method S, $\boldsymbol{\gamma}_S$ (a vector of length I_S, the number of groups observed by method S). We drop the j and t subscript because this mapping does not depend on the identity of the community being sampled. Then,

$$\boldsymbol{\gamma}_S = \mathbf{M}_{MS}\boldsymbol{\gamma}_M \tag{9}$$

$\mathbf{M}_{MS}$ is a I_S by I_M matrix.

For example, if there are four species in the true community and method S only observes three groups, the matrix $\mathbf{M}_{MS}$ could look like this

$$\mathbf{M}_{MS} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{10}$$

This might happen if species 2 and species 4 (columns 2 and 4, respectively) were from the same genus and the PCR primer from method S can only resolve those two species at the genus level. To provide a further

example, take an invented community of four species with $\boldsymbol{\gamma}_M = \{1, 15, 6, 7\}$ individuals in the community. The true community as observed through method S would be

$$\boldsymbol{\gamma}_S = \mathbf{M}_{MS}\boldsymbol{\gamma}_M \tag{11}$$

$$\begin{bmatrix} 1 \\ 22 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 15 \\ 6 \\ 7 \end{bmatrix} \tag{12}$$

and so $\boldsymbol{\gamma}_S$ is a linear combination of the true community. Of course there is no requirement that elements of $\boldsymbol{\gamma_M}$ be integers, but that makes the above example easy and transparent.

It is easy to incorporate this added complexity into the models in the previous section. If we assign method S to be manual counts and W to be the Mifish PCR primer, we need to construct a main list of species to define $\boldsymbol{\gamma}_M$ and build two mapping matrices, $\mathbf{M}_{MS}$ and $\mathbf{M}_{MW}$ that determine which species or species-groups are observed by each method. We can then add an additional subscript for each additional method and use the same form as above. For example,

$$\log(\theta_{Sijt}) = \log(\gamma_{Sijt}) + \log(P_{Sjt}) + \log(V_{Sjt}) \tag{13}$$

$$\log(\lambda_{Wijtk}) = \log(\beta_{Wijt}) + N_{W,PCR}\log(1 + \alpha_{Wi}) + \eta_{Wjtk} \tag{14}$$

And with additional sampling methods, we can make different mappings from the true abundance to the observations of each method.

Estimation and Identifiability

We developed and fit the above model in a Bayesian framework using the Stan language, as implemented in *RStan* (Stan Development Team 2020). All code is available as supplementary material. Table 1 provide prior distributions used in the model.

We ran five MCMC chains with 1,000 warmup and 4,000 sampling iterations. We retained every other MCMC sample. We initiated each chain at randomly determined starting values. The model converged ($\hat{R} < 1.01$; Gelman-Rubin diagnostics) and had no divergent transitions. We performed standard posterior predictive checks to assess model fit.

Table 1: Prior and parameter descriptions for the Stan Model.

Parameter & Prior	Description
$\alpha_i \sim Beta(1,1)$	Amplification efficiency for species i
$\log(\gamma_{Mijt}) \sim Normal(0,4)$	True biomass of each species at each site-year
$\log(\eta_{jtk}) \sim Normal(-4,4)$	Estimated offset for each PCR reaction at each site-year
$\tau_0 \sim Normal(0,2)$	Negative Binomial shape parameter intercept
$\tau_1 \sim Normal(0,2)$	Negative Binomial shape parameter slope

References

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/.