Notes on RAD MS

Species abundances in a community are derived from a few basic processes: (1) species’ interactions with their environment, (2) species’ interactions with other organisms, and (3) migration of individuals into and out of the area of interest.

Hughes 1986 demonstrated the usefulness of a dynamics-based approach, finding that a large proportion of sampled communities display abundance curves that were qualitatively similar to those produced by his relatively simple dynamic model.

From Hughes 1986:

*The dynamics model predicts the species-abundance pattern of most samples with greater accuracy and provides an explanation of species abundance based on recognized and testable ecological principles.*

Using Lotka-Volterra (LV) models allow us to address the first two of these issues (for our purposes we ignore migration, although it could conceivably be incorporated into the model). In this model species interactions with each other are explicit, and their interaction with the environment is implied through the implementation of growth rates, carrying capacity, and “self” interactions.

An important distinction between this study of abundance distributions and those of others is that we explicitly look at communities of all interaction types, while most studies of abundance distributions have focused on horizontal communities only (e.g., all plants in a given area). The restriction to horizontal communities limits the potential influence of species interactions, as these communities typically only exhibit competitive or facilitative interactions.

We seek to answer three main questions about the rank-abundance distribution (RAD) using simulations based on the LV model.

1. How does variation in these species interaction drive variation in the shape of their abundance distribution?
2. Can the LV model generate communities with species RAD that closely mirror those of observed communities?

To address the first question, we will use linear regression to determine how the number and strength of different species interaction distributions in the community affect the fitted RAD. We also include species’ growth rates, self-interactions, and total number of species in our regression. The second question will be addressed using classification and regression tree analysis, paired with random forest models. These methods will highlight the importance of the different predictor variables in generating communities that lie either in or out of the parameter range of naturally occurring communities.

The Zipf model has been shown to fit reasonably well to many species abundance distributions (but see Baldridge for otherwise), and may easily be used in the rank-abundance form using the `sads` R package. This RAD model has a single parameter *s* that describes the shape of the distribution, where lower values indicate more evenly distributed abundances.

We compiled human gut microbiome data from the Human Microbiome Project (HMP), as well as time-series relative abundances from four sampled individuals (time-series lengths range from ~250 days to 360). To standardize the samples we resampled from the original data 2000 individuals. Any HMP samples that did not initially have at least 2000 reads were excluded.

The resampled communities were fit with the Zipf model.

Our regression explains a significant amount of the variation in fitted *s* values to simulated data.

A few reasons why simulated distributions do not match observed distributions are that: (1) observed communities are likely an amalgam of several spatially structured communities throughout the gut because of the sampling method, (2) migration into and out of human gut microbial communities, (3) copy number variation in the 16S RNA gene may be skewing results, (4) current methods do not adequately sample the gut community, or (5) clustering of OTUs does not accurately represent the true species distribution. On the other hand, a number of assumptions inherent in the simulation process may also be to blame. For example, we simulated initially randomly interacting communities while most real communities in macrobiological systems are decidedly non-random. We also do not know the true distribution of interactions or their strengths in real microbial communities, so we may have chosen incorrect distributions for these parameters, or not adequately sampled parameter space.

We have also likely not captured the full extent of the influence the environment may have on the abundance distribution. Without explicitly modeling this interaction (e.g., through consumption of allochthonous resources), we are limited in our ability to determine the effect it has. The species’ growth rate and self-interaction term may not be adequate proxies for their interaction with the environment.