Using a co-occurrance network

Jim Brunner

June 19, 2017

We are looking at creating ecological networks of a microbiome. Right now, I have built two networks, in the form of graphs $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$. In both graphs, vertex are labeled by taxa name. I'm going to conflate the vertices and their label. The edge sets are defined by co-incidence and co-occurence, respectively. Given a set of samples with abundancs of organisms, we map the abundances to discrete levels, as proportions of the maximum abundance *of that organism*. Precisely, let the samples be s_i and raw abundance of organism j in sample i be r_{ji}. I map the abundances according to

$$a(r_{ji}) = \begin{cases} \lfloor \left(\frac{r_{ji}}{\max_{s_k}(r_{jk})}\right) n \rfloor + 1 & \frac{r_{ji}}{\max_{s_k}(r_{jk})} \geq m \\ 0 & \frac{r_{ji}}{\max_{s_k}(r_{jk})} < m \end{cases}$$

where m is some minimum. Then, I create weighted edges between vertices (where 0 weight means no edge exists) where the weights of edges in $\mathcal{E}_1$ are

$$w^1_{jk} = \frac{\|\{i : a(r_{ji}) = a(r_{ki}) \neq 0\}\|}{S}$$

where S is the total number of samples. That is, we count the propotion of samples in which the two taxa appear at the same discritized level.

The second network accounts for random coincidence of taxa in a sample, following [1]. It begins with $\mathcal{G}_1$, and compares to a null model N. The null model is defined in the following way.

$$A_j = \sum_{s_i} \mathbf{1}_{a(r_{ji}) \neq 0}$$

and

$$S_i^l = \sum_{v_j} \mathbf{1}_{a(r_{ji}) = l}$$

Then, N assumes that if

$$X_{jil} \sim binom\left(A_j, \frac{S_i^l}{\sum_{il} S_i^l}\right)$$

then $P(a^N_{ji} = l) = 1 - P(X_{jil} = 0)$. This allows us to calculate the probability of co-incidence of taxa under the null model. Let w^N_{jk} be

$$w^N_{jk} = \|\{i : a^N_{ji} = a^N_{ki} \neq 0\}\|$$

This is the similar to the co-incidence model but now randomized. Then,

$$P(w^N_{jk} = K) = \sum_{\{A \subset \mathcal{V}_2 : |A| = k\}} \prod_{l \in A} a_{jl} a_{kl} \prod_{l \notin A} (1 - a_{jl} a_{kl})$$

Ideally, we would then define $\mathcal{E}_2$ by the weights

$$w^2_{jk} = \begin{cases} 1 & P(w^N_{jk} \geq w^1_{jk}) \geq t \\ 0 & P(w^N_{jk} < w^1_{jk}) > t \end{cases}$$

However, that probability is intractible to compute. Instead, we take

$$\widetilde{P}(w^N_{jk} = K) = \sum_{l=0}^{i} \binom{N_1}{l} \binom{N_2}{K-l} p_1^j p_2^{K-l} (1-p_1)^{N-l} (1-p_2)^{N-K+l}$$

where

$$p_1 = p_a - \left(\frac{N_2}{N_1}\frac{N(\mu - \sigma^2) - \mu^2}{N^2}\right)^{1/2}$$

and

$$p_2 = p_a - \left(\frac{N_1}{N_2}\frac{N(\mu - \sigma^2) - \mu^2}{N^2}\right)^{1/2}$$

Finally, N_1, N_2 are to ensure that $p_1, p_2 \in [0,1]$. It turns out we need:

$$\frac{\mu N(1-p_a) - N\sigma^2}{N(1-p_a) - \sigma^2} \leq N_2 \leq \frac{\mu^2}{\mu - \sigma^2}$$

with μ the mean of the real distrubution, σ^2 the variance, and $p_a = \frac{1}{S}\sum_i a_{ji}a_{ki}$. So, we take

$$w^2_{jk} = \begin{cases} 1 & \widetilde{P}(w^N_{jk} \geq w^1_{jk}) \geq t \\ 0 & \widetilde{P}(w^N_{jk} < w^1_{jk}) > t \end{cases}$$

The question now is what can we do these networks?

First, assesing GOTTCHA reads. A single GOTTCHA read would produce a network with each connected component complete. I think the co-incidence network is more appropriate. We want to assess the probability that you see this set together. We have the probability that you see any vertex pair (estimated) as the edge weights of $\mathcal{G}_1$. Precisely, the edge weights are

$$w^1_{jk} = P(j \,\&\, k \in S^l_i)$$

where S^l_i is sample i at discrete abundance level l. It might be useful to have the directed weight graph where

$$w^3_{jk} = P(j \in S^l_i | k \in S^l_i)$$

but that wouldn't be hard, because then

$$w^3_{jk} = S\frac{w^1_{jk}}{\|\{i : a(r_{ki}) > 0\}\|}$$

Anyway, let's start with the simplest case of one abundance level. Assume GOTTCHA found taxa $a, b, c, ..., n$. Maybe the first thing we would want is

$$P(a|b,c,d,...,n),\ P(b|a,c,d,...,n),\ etc$$

Clearly, we can see directly $P(a|b)$, etc. We can also get a bound for triplets (assuming $P(c,b) \neq 0$):

$$P(a|b,c) \leq \frac{\min_{(i,j)\subset\{a,b,c\}}(P(i,j))}{P(b,c)}$$

but we can't do any better than triplets explicitly, because we don't have any sort of independence (conditional or otherwise) and because our network is not acyclic.

We can probably learn something from asking about the connectivity of the induced subgraph. Notice that if it isn't complete, then one of the $P(a,b)$ is 0 above.

What does the connectivity of the induced subgraph of $\mathcal{G}_2$ tell us? If it is connected, that's good.

References

[1] Heiko Mller and Francesco Mancuso. Identification and analysis of co-occurrence networks with netcutter. *PLOS ONE*, 3(9):1–16, 09 2008.