A Poisson-multivariate normal hierarchical model for measuring microbial conditional independence networks from metagenomic count data

Surojit Biswas[1], Derek S. Lundberg[2], Jeffery L. Dangl[2,3,4,5], Vladimir Jojic[6]

[1] Department of Statistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[2] Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[3] Howard Hughes Medical Institute, University of North Carolina, Chapel Hill, NC, 27599, USA
[4] Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC, 27599, USA
[5] Department of Immunology, University of North Carolina, Chapel Hill, NC, 27599, USA
[6] Department of Computer Science, University of North Carolina, Chapel Hill, NC, 27599, USA

Introduction

Microbes are the most diverse form of life on the planet. Many associate with higher eukaryotes, including humans and plants, and perform key metabolic functions that underpin host viability [1, 2]. Importantly, they coexist in these ecologies in various symbiotic relationships [3]. Understanding the structure of their interaction networks may simplify the list of microbial targets that can be modulated for host benefit.

Microbiomes can be measured by sequencing all host-associated 16S rRNA gene content. Because the 16S gene is a faithful phylogenetic marker, this approach readily reveals the taxonomic composition of the host metagenome [4]. Given such sequencing experiments output an integral, non-negative number of sequencing reads, the final output for such an experiment can be summarized in a n-samples $\times$ s-taxa count table, Y, where Y_{ij} denotes the number of reads that map taxon j in sample i. It is assumed Y_{ij} is proportional to taxon j's true abundance in sample i.

To study interrelationships between taxa, we require a method that transforms Y into an undirected graph represented by a symmetric and weighted $s \times s$ adjacency matrix, A, where a non-zero entry in position (i, j) indicates an association between taxon i and taxon j. Correlation-based methods are a popular approach to achieve this end [5–7]. Nevertheless, correlated taxa need not directly interact if, for example, they are co-regulated by a third taxon. Gaussian graphical models remedy this concern by estimating a conditional independence network in which $A_{ij} = 0$ if and only if taxon i and taxon j are conditionally independent given all remaining taxa under consideration [8–10]. However, they also assume the columns of Y are normally distributed, which is unreasonable for a metagenomic sequencing experiment. Finally, neither correlation nor Gaussian graphical modeling offer a systematic way to control for confounding predictors, such as measured biological covariates (e.g. body site, or plant fraction), experimental replicate, sequencing plate, or sequencing depth.

In this work, we develop a Poisson-multivariate normal hierarchical model that can account for correlation structure among count-based random variables. Our model controls for confounding predictors

at the Poisson layer, and captures direct taxon-taxon interactions at the multivariate normal layer using an L_1 penalized precision matrix.

Methods

Preliminaries

Let n, p, and s denote the number of samples, number of predictors, and the number of taxa under consideration, respectively. Let Y be the $n \times s$ response matrix, where Y_{ij} denotes the count of taxon j in sample i. Let X be the $n \times p$ design matrix, where X_{ij} denotes predictor j's value for sample i. For a matrix M, we will use the notation $M_{:i}$ and $M_{i:}$ to index the entire i^{th} column and row, respectively.

The model

We wish to model conditional independence relationships among bacterial taxa measured in a metagenomic sequencing experiment while also controlling for the confounding predictors encoded in X. Toward this end, we propose the following Poisson-multivariate normal hierarchical model.

$$\begin{aligned} w &\sim \text{Multivariate-Normal}\left(\mu, \Sigma^{-1}\right) \\ Y &\sim \text{Poisson}(\exp\{X\beta + w\}) \end{aligned}$$

Here μ and Σ^{-1} are the $1 \times o$ mean vector and $o \times o$ precision matrix of the multivariate normal, and w is an $n \times o$ latent abundance matrix. The coefficient matrix, β, is $p \times o$ such that β_{ij} denotes predictor i's coefficient for taxon j.

The likelihood of this model is given by

$$\sum_{j=1}^{o}\sum_{i=1}^{n}\left[y_{ij}(x_{i:}\beta_{:j} + w_{ij}) - \exp\{x_{i:}\beta_{:j} + w_{ij}\}\right] + \frac{n}{2}\log|\Sigma^{-1}| - \frac{n}{2}\text{tr}\left(\hat{\Sigma}(w)\Sigma^{-1}\right)$$

where $\Sigma(w)$ is the empirical covariance matrix of w.

Intuitively, the columns of w are adjusted, "residual" abundance measurements of each taxa, after controlling for confounding predictors in X. Therefore, we wish to model conditional independence at the level of these latent abundances, rather than the observed counts. Recall if $\Sigma^{-1}_{ij} = 0$, then $w_{:i}$ and $w_{:j}$ are conditionally independent. To improve conditioning and the saliency of the result, we therefore impose an adjustable L_1-penalty on the entries of the precision matrix during optimization.

Model learning

The L_1-penalized likelihood, modulo unnecessary constants, is given by

$$\underset{\beta,w,\mu,\Sigma^{-1}}{\text{argmax}}\sum_{j=1}^{o}\sum_{i=1}^{n}\left[y_{ij}(x_{i:}\beta_{:j} + w_{ij}) - \exp\{x_{i:}\beta_{:j} + w_{ij}\}\right] + \frac{n}{2}\log|\Sigma^{-1}| - \frac{n}{2}\text{tr}\left(\hat{\Sigma}(w)\Sigma^{-1}\right) - \frac{\lambda n}{2}||\Sigma^{-1}||_1$$

where λ is a tuning parameter, and $||\cdot||_1$ denotes the L_1-norm.

We optimize this objective using an iterative conditional modes algorithm in which parameters are sequentially updated to their mode value given current estimates of the remaining parameters [11]. Given an estimates of w, μ, and Σ^{-1}, the conditional objective for β is given by,

$$\underset{\beta}{\operatorname{argmax}} \sum_{j=1}^{o} \sum_{i=1}^{n} [y_{ij}(x_{i:}\beta_{:j} + \hat{w}_{ij}) - \exp\{x_{i:}\beta_{:j} + \hat{w}_{ij}\}]$$

This is efficiently and uniquely optimized by setting $\beta_{:k}$ to the solution of the Poisson regression of $Y_{:k}$ onto X using $w_{:k}$ as an offset, for all $k \in \{1, 2, \ldots, o\}$.

Given estimates for β, Σ^{-1} and μ, the conditional objective for w is given by

$$\underset{w}{\operatorname{argmax}} \sum_{j=1}^{o} \sum_{i=1}^{n} \left[y_{ij}w_{ij} - \exp\{x_{i:}\hat{\beta}_{:j} + w_{ij}\} \right] - \frac{n}{2} \operatorname{tr}\left(\hat{\Sigma}(w)\hat{\Sigma}^{-1} \right)$$

Each row of w is independent of all other rows in this objective and can therefore be updated separately. To obtain the conditional update for $w_{i:}$, we apply Newton-Raphson. The gradient vector, g_i, and Hessian, H_i, are given by

$$g_i = y_{i:} - \exp\{x_{i:}\hat{\beta}_{:j} + w_{ij}\} - (w_{i:} - \hat{\mu})\hat{\Sigma}^{-1} \quad H_i = -\hat{\Sigma}^{-1} - \operatorname{diag}(\exp\{x_{i:}\hat{\beta} + w_{i:}\})$$

Because $\hat{\Sigma}^{-1}$ is positive-definite and $\exp\{x_{i:}\hat{\beta} + w_{i:}\} > 0$ for all components, H_i is always negative-definite. Thus, the conditional update for $w_{i:}$ is a unique solution.

Given β, Σ^{-1} and w, the conditional objective for μ is maximized by taking the sample mean of each column of w.

Given β, w, and μ, the conditional objective for Σ^{-1} is given by,

$$\underset{\Sigma^{-1}}{\operatorname{argmax}} \log|\Sigma^{-1}| - \operatorname{tr}\left(\hat{\Sigma}(\hat{w})\Sigma^{-1} \right) - \lambda||\Sigma^{-1}||_1$$

which is efficiently optimized using the graphical lasso [9].

Synthetic Community Validation Experiment

To test the model with real data, we constructed a 9 member artificial community composed of *Escherichia coli* (negative colonization control) and 8 other bacterial strains originally isolated from plant roots grown in two wild soils [2]. For each of 46 sterile plants, we inoculated the 9 isolates in varying relative abundances in order to perturb their underlying interaction structure (Figure 1a). More specifically, for approximately half of the plants, one or more of the strains were randomly dropped out, and for the other half, all strains were present, but ranged in input abundance from 0.5-50%.

Plants were grown in an inert and sterile calcine-clay soil. Plant roots were harvested 4 weeks post inoculation, and the abundance of root-associated isolates was measured with 16S rRNA profiling of the V4 variable region using the method described in [12]. All consensus sequences (ConSeqs) – sequencing reads adjusted for PCR-amplification bias [12] – for each root sample were mapped using the Burrows Wheeler Aligner [13], to a previously constructed sequence database of each isolate's V4 region. Mapped ConSeqs to a given isolate in a given sample were counted and subsequently assembled into a 46-samples $\times$ 9-isolates count matrix, visualized in Figure 1b.

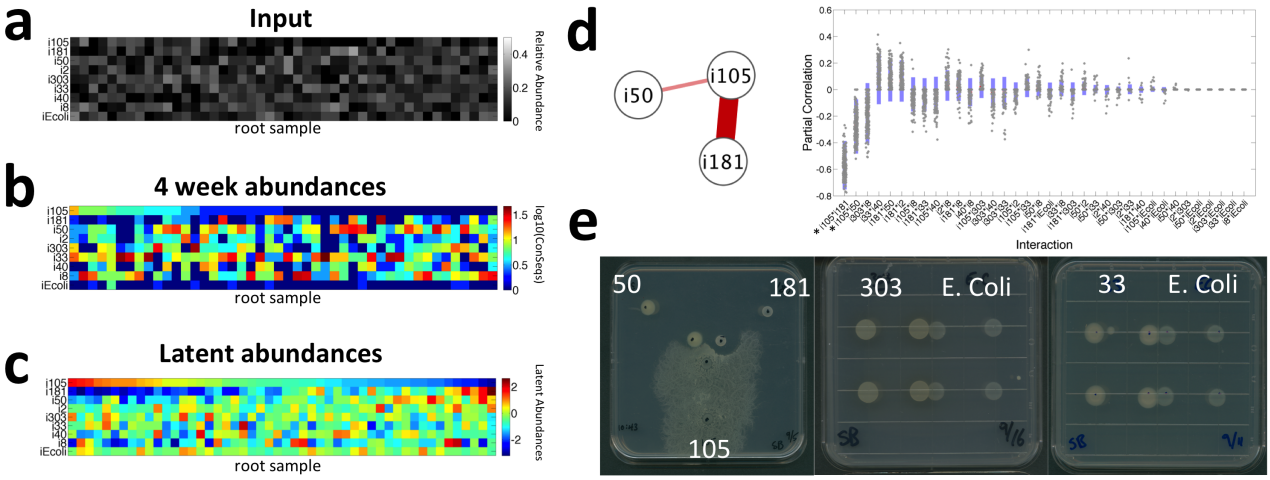

Figure 1: **Re-colonization and isolate-isolate interaction results from the 9 member synthetic community.** a) Isolates $\times$ plants (samples) input abundance matrix. b) $\log_{10}$ transformed raw abundances of each isolate obtained at the time of harvest (4 weeks post inoculation). Abundances are measured in ConSeqs [12], which are PCR bias corrected sequencing reads. c) Inferred latent abundance matrix. d) Network visualization of significant interactions (left) and bootstrap sampling distributions of all entries in the precision matrix (right). An interaction was considered significant if the 5 and 95 percentile interval of its bootstrap distribution did not contain 0. e) Co-plating experiments to test model predictions of interactions and non-interactions. The left-most plate tests the network illustrated in (d), and the middle and right plate test interaction predictions that were not significant. The images shown are representative of three independent co-platings.

Results

We applied our model to the 46 root-samples $\times$ 9 isolates count matrix. Starting input abundances (Figure 1a), $\log$-sequencing depth, and harvester were entered as predictors. To assess the significance of interaction predictions, the model was bootstrapped 200 times.

Figure 1c illustrates the inferred latent abundance matrix, $\hat{w}$. Clear negative correlation can be observed between isolates 105 and 181 and to a smaller extent between isolates 105 and 50. The precision matrix that supports these interactions suggests 50 and 181 are conditionally independent given 105, and that interactions between (105,181) and (105,50) are inhibitory in nature.

In vitro co-plating experiments corroborate the model's predictions exactly in direction and also semi-quantitatively (Figure 1e). In particular, they show that (105, 181) and (105, 50) are, indeed, antagonistic interaction pairs, and moreover, that 181 and 50 are the inhibitors. Additionally, the (181, 105) inhibition appears more pronounced than the (181, 50) inhibition, just as the model suggests. The model also predicts conditional independence of 50 and 181 given 105. Indeed, the inward facing edges of the 181 and 105 colonies do not appear deformed, and therefore suggest a non-interaction. Co-platings of the (303, *E. coli*) and (33, *E. coli*) interaction pairs reveal no significant interaction and thereby support the model's predictions of non-interaction for these pairs.

Discussion

We demonstrated our Poisson-multivariate normal hierarchical model can infer true, direct microbe-microbe interactions in real data. Though not illustrated for brevity, we find that proper modeling

of confounding predictors is necessary to detect the (105, 181) and (105, 50) interactions. Without controlling for starting input abundances, harvester, and particularly $\log$-sequencing depth, the model does not detect any significant interactions. In simulation experiments (not shown), we find that our model consistently maintains a lower false discovery rate than Spearman correlation, a sparse correlation threshold method, SparCC [6], and the Nonparanormal SKEPTIC – a semi-parametric version of the graphical lasso [14] – when there are confounding predictors. Without such covariates, our model and the Nonparanormal SKEPTIC perform comparably in the number of falsely discovered edges, and both handily outperform the correlation based methods. Finally, we have also extended our method to model multivariate, count-based time-series. We are currently testing it using 20-member synthetic community perturbation experiments that are harvested over 8 time points, 4-7 weeks post inoculation.

References

[1] Human Microbiome Project Consortium The. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–14, June 2012.

[2] Derek S. Lundberg, Sarah L. Lebeis, Sur Herrera Paredes, Scott Yourstone, Jase Gehring, Stephanie Malfatti, Julien Tremblay, Anna Engelbrektson, Victor Kunin, Tijana Glavina Del Rio, Robert C. Edgar, Thilo Eickhorst, Ruth E. Ley, Philip Hugenholtz, Susannah Green Tringe, and Jeffery L. Dangl. Defining the core Arabidopsis thaliana root microbiome. *Nature*, 488(7409):86–90, August 2012.

[3] Allan Konopka. What is microbial community ecology? *The ISME journal*, 3(11):1223–30, November 2009.

[4] Nicola Segata, Daniela Boernigen, Timothy L Tickle, Xochitl C Morgan, Wendy S Garrett, and Curtis Huttenhower. Computational meta'omics for microbial community studies. *Molecular systems biology*, 9(666):666, January 2013.

[5] Karoline Faust, J Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology*, 8(7):e1002606, January 2012.

[6] Jonathan Friedman and Eric J Alm. Inferring Correlation Networks from Genomic Survey Data. *PLoS computational biology*, 8(9):1–11, 2012.

[7] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature reviews. Microbiology*, 10(8):538–50, August 2012.

[8] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, June 2006.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, 9(3):432–41, July 2007.

[10] Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Found. Trends Mach. Learn.*, 1(1935-8237):1–305, 2008.

[11] Julian Besag. On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.

[12] Derek S Lundberg, Scott Yourstone, Piotr Mieczkowski, Corbin D Jones, and Jeffery L Dangl. Practical innovations for high-throughput amplicon sequencing. *Nature methods*, 10(10):999–1002, October 2013.

[13] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–95, March 2010.

[14] John Lafferty and Larry Wasserman. The Nonparanormal skeptic. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.