Robust and Scalable Models of Microbiome Dynamics

Travis E. Gibson[1]
tgibson@mit.edu

Georg K. Gerber[1,2]
ggerber@bwh.harvard.edu

[1]Massachusetts Host Microbiome Center
Brigham and Women's Hospital & Harvard Medical School

[2]Harvard-MIT Health Sciences and Technology

35th International Conference on Machine Learning
July 12th, 2018

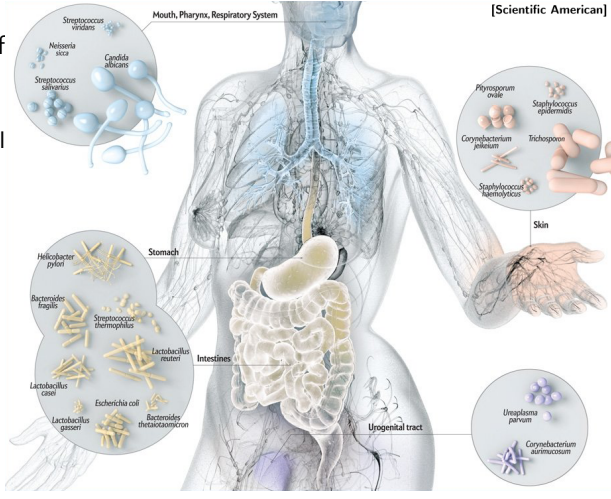
Poster #2

Outline

1. What is the microbiome
2. From sequences to reads to microbial interactions
3. Bayesian nonparametric model for microbe dynamics

The Microbiome

1. The **microbiome** is the aggregate of microorganisms that resides on or within any of a number of human tissues and biofluids:
   - skin, mammary glands, placenta, seminal fluid, uterus, ovarian follicles, lung, saliva, oral mucosa, conjunctiva, biliary and **gastrointestinal tracts**) [wikipedia]
2. 10^{14} Microbes in/on your body [Sender et al. *PLoS Biology* 2016]
3. 3.3 million genes compared to 23,000 human genes [Qin et al. *Nature* 2010]
4. Play a role in a variety of human diseases:
   - infections, arthritis, food allergy, cancer, inflammatory bowel disease, neurological diseases, and obesity/diabetes



Bacteriotherapy

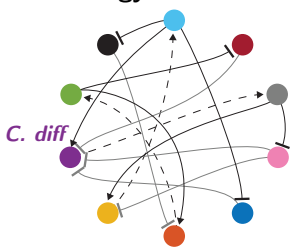
Bacteriotherapy: communities of bacteria administered to patients for specific therapeutic applications

- **"bugs-as-drugs"**

Clostridium difficile infection

- Causes serious diarrhea (14K deaths/yr)
- Antibiotics disrupt helpful bacteria in gut
- Increasingly difficult to treat with conventional therapies (more antibiotics): 20-30% recurrence rate

Pharmacology meets Ecology



microbial interaction network

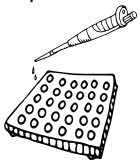
positive microbe A produces a small molecule (metabolite) that microbe B needs

negative two microbes competing for the same niche

what if there were 300 bugs in the network?

Workflow in our lab

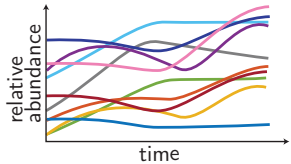
batch experiments



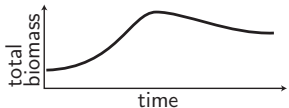
chemostat

animal experiments

- 16S rRNA on MiSeq (reads) for relative abundances of species

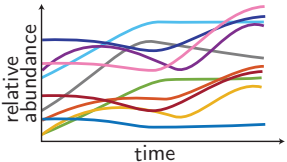


- 16S rRNA qPCR (universal primers) for bacterial biomass



Sequencing to obtain microbe relative abundances

- **16S rRNA on MiSeq (reads) for relative abundances of species**



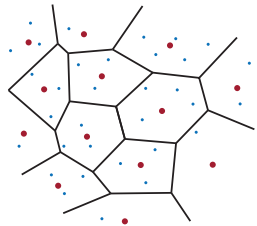
1. Fastq file

```
@sequence-id
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAA9#:<#<;<<<????#=
```

nucleobase sequence

quality scores

2. Sequences are clustered



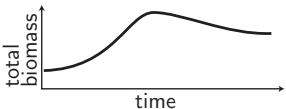
3. Read Table

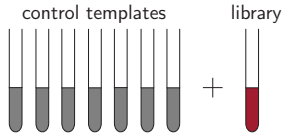
	sample1	sample2	...
bacteria1	11	1004	
bacteria2	0	7	
bacteria3	301	275	
⋮			

Reads $\sim$ Negative Binomial

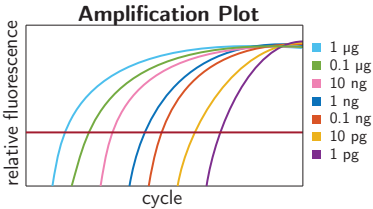
Quantitative PCR for total bacterial biomass

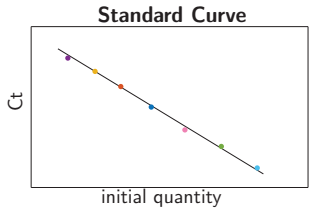
- **16S rRNA qPCR (universal primers) for bacterial biomass**



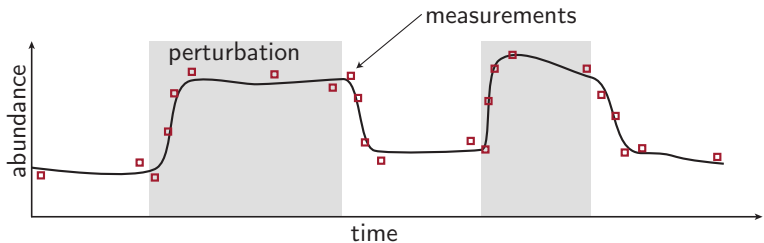


Dilute control template and library for quantification





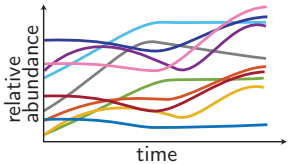
Irregular and sparse measurements



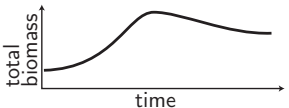
- **measurements - irregular, sparse & noisy**

Learning microbial interaction networks

- 16S rRNA on MiSeq (reads) for relative abundances of species



- 16S rRNA qPCR (universal primers) for bacterial biomass



- measurements - irregular, sparse & noisy

Interaction Network

Abundance of microbe i at time t : $\mathbf{x}_{t,i}$

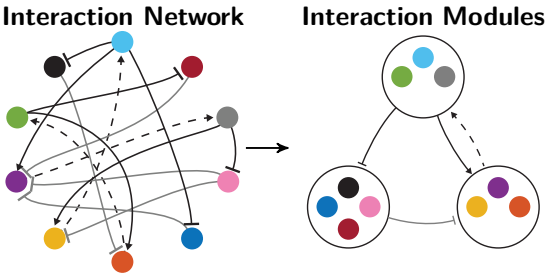
$$\frac{\mathrm{d}\mathbf{x}_{t,i}}{\mathrm{d}t} = \boldsymbol{\alpha}_i \mathbf{x}_{t,i} + \boldsymbol{\beta}_{ii} \mathbf{x}_{t,i}^2 + \sum_{j \neq i} \boldsymbol{\beta}_{ij} \mathbf{x}_{t,i} \mathbf{x}_{t,j}$$

growth, self limiting, interaction

Previous literature specific to the microbiome

- Log transform dynamics $\rightarrow$ Linear Regression + L2 [Stein et al. *PLoS Comput Biology* 2013]
- Sparse linear regression with bootstrap aggregation [Fisher et al. *PLoS One* 2014]
- Bayesian model with deterministic dynamics (independent filtering) [Bucci et al. *Genome Biology* 2016]
- Extended Kalman Filter [Alshawaqfeh et al. *BMC Genomics* 2017]

Goal with our model and short literature review



- 300 species
- 90,000 interactions
- Redundant gene function

Three main contributions in our model

1. Clustering (interaction modules)
   - Dirichlet Process (DP)
     [Rasmussen, *Advances in Information Processing Systems* 2000]
     [Neal, *Journal of computational and graphical statistics*, 2000]
2. Edge selection (structure learning, variable selection)
   - Bayesian Networks
     [George and McCulloch, *Journal of the ASA*, 1993]
     [Heckerman, *A Tutorial on Learning with Bayesian Networks*, 2008]
3. Introduction of an auxiliary variable between the measurement model and latent state

Back to the basic time series model

- Abundance of microbe i at time t : $\mathbf{x}_{t,i}$

$$\frac{\mathrm{d}\mathbf{x}_{t,i}}{\mathrm{d}t} = \boldsymbol{\alpha}_i \mathbf{x}_{t,i} + \boldsymbol{\beta}_{ii} \mathbf{x}_{t,i}^2 + \sum_{j \neq i} \boldsymbol{\beta}_{ij} \mathbf{x}_{t,i} \mathbf{x}_{t,j} + \frac{\mathrm{d}\mathbf{w}_{t,i}}{\mathrm{d}t}$$

growth, self limiting, interaction, stochastic disturbance

- Convert to discrete time

$$\mathbf{x}_{k+1,i} = \mathbf{x}_{k,i} + \left(\boldsymbol{\alpha}_i \mathbf{x}_{k,i} + \boldsymbol{\beta}_{ii} \mathbf{x}_{k,i}^2 + \sum_{j \neq i} \boldsymbol{\beta}_{ij} \mathbf{x}_{k,i} \mathbf{x}_{k,j} \right) \Delta_k + (\mathbf{w}_{k+1,i} - \mathbf{w}_{k,i}) \sqrt{\Delta_k}$$

discrete time step size

Next we discuss the three main ingredients to our model

1. Clustering (interaction modules)
2. Edge selection (structure learning, variable selection)
3. Introduction of an auxiliary variable between the measurement model

Complete Model

Dirichlet Process

$$\boldsymbol{\pi}_{\mathbf{c}} \mid \boldsymbol{\alpha} \sim \texttt{Stick}(\boldsymbol{\alpha})$$

$$\mathbf{c}_i \mid \boldsymbol{\pi}_{\mathbf{c}} \sim \texttt{Multinomial}(\boldsymbol{\pi}_{\mathbf{c}})$$

$$\mathbf{b}_{\mathbf{c}_i,\mathbf{c}_j} \mid \boldsymbol{\sigma}_{\mathbf{b}} \sim \texttt{Normal}(0, \boldsymbol{\sigma}_{\mathbf{b}}^2)$$

Edge Selection (Structure)

$$\mathbf{z}_{\mathbf{c}_i,\mathbf{c}_j} \mid \boldsymbol{\pi}_{\mathbf{z}} \sim \texttt{Bernouli}(\boldsymbol{\pi}_{\mathbf{z}})$$

Self Interactions

$$\mathbf{a}_{i,1}, \mathbf{a}_{i,2} \mid \boldsymbol{\sigma}_{\mathbf{a}} \sim \texttt{Normal}(0, \boldsymbol{\sigma}_{\mathbf{a}}^2)$$

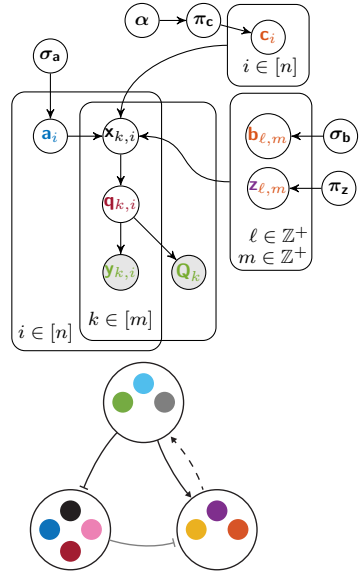
Dynamics

$$\mathbf{x}_{k+1,i} \mid \mathbf{x}_k, \mathbf{a}_i, \mathbf{b}, \mathbf{c}, \mathbf{z}, \boldsymbol{\sigma}_{\mathbf{w}} \sim$$

$$\texttt{Normal}\left(\mathbf{x}_{k,i} + \mathbf{x}_{k,i}\left(\mathbf{a}_{i,1} + \mathbf{a}_{i,2}\mathbf{x}_{k,i} + \sum_{\mathbf{c}_j \neq \mathbf{c}_i} \mathbf{b}_{\mathbf{c}_i,\mathbf{c}_j}\mathbf{z}_{\mathbf{c}_i,\mathbf{c}_j}\mathbf{x}_{k,j}\right), \Delta_k \boldsymbol{\sigma}_{\mathbf{w}}^2\right)$$

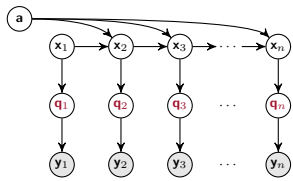
Constraint and Measurement Model

aux	$\mathbf{q}_{k,i} \mid \mathbf{x}_{k,i} \sim \texttt{Normal}(\mathbf{x}_{k,i}, \boldsymbol{\sigma}_{\mathbf{q}}^2)$
reads	$\mathbf{y}_{k,i} \mid \mathbf{q}_{k,i} \sim \texttt{NegBin}(\phi(\mathbf{q}_k), \epsilon(\mathbf{q}_k))$
qPCR	$\mathbf{Q}_k \mid \mathbf{q}_{k,i} \sim \texttt{Normal}\left(\sum_i \mathbf{q}_{k,i}, \sigma_{\mathbf{Q}_k}^2\right)$



Simplified model unraveled in time - auxiliary variable

$$\mathbf{x}_{t+1,i} \mid \mathbf{x}_t, \mathbf{a} \sim \texttt{Normal}(\mathbf{a}_i^\top f(\mathbf{x}_t), \sigma^2_{\mathbf{x}_i})$$
$$\mathbf{q}_{k,i} \mid \mathbf{x}_{k,i} \sim \texttt{Normal}(\mathbf{x}_{k,i}, \boldsymbol{\sigma}^2_{\mathbf{q}})$$
$$\mathbf{q}_{k,i} \sim \texttt{Uniform}[0, L)$$
$$\mathbf{y}_{k,i} \mid \boldsymbol{\sigma}_{\mathbf{y}}, \mathbf{q}_{k,i} \sim \texttt{Normal}_{\geq 0}(\mathbf{q}_{k,i}, \boldsymbol{\sigma}^2_{\mathbf{y}})$$
$$\mathbf{a}_i \sim \texttt{Normal}(0, \sigma^2_{\mathbf{a}_i})$$



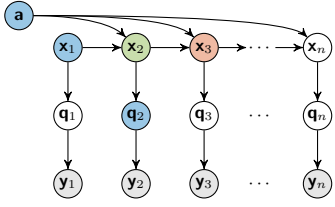
Prior on **q** is positive, relaxing the distribution on the dynamics for **x**

Parameter inference Gibbs step: $\mathbf{a}^{(g+1)} \sim p_{\mathbf{a}|\mathbf{x}}(\cdot \mid \mathbf{x}^{(g)})$

- Direct sampling from the posterior possible (Bayesian Regression!)

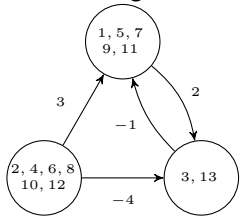
Sampling for other variables

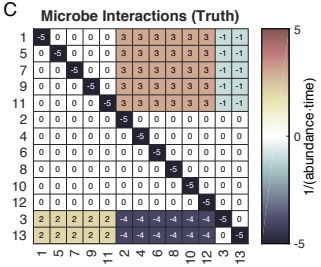
- Collapsed Gibbs sampling for Dirichlet Process and Edge Selection (integrate out **a**)
- Filtering is still challenging but easy to design proposals for (MH)

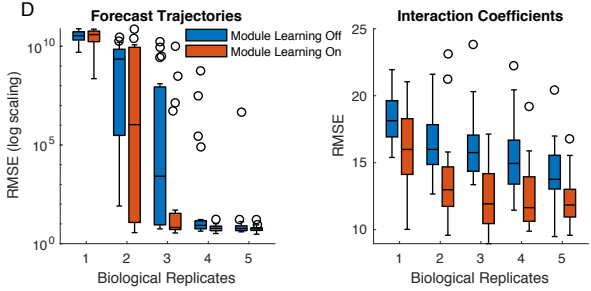


Synthetic experiment

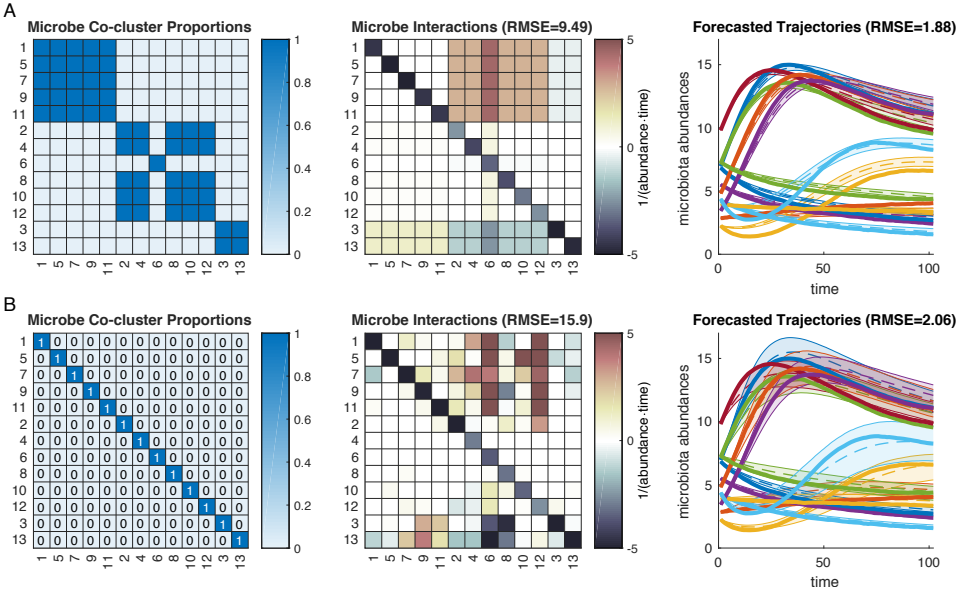
- Comparing inference with and without clustering enabled
- Ground truth interaction network





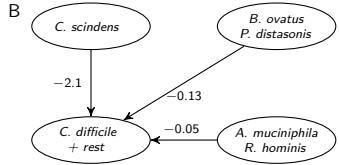


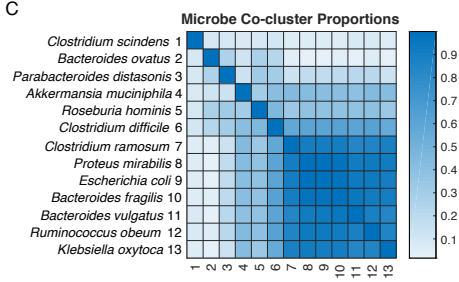
Synthetic experiment continued

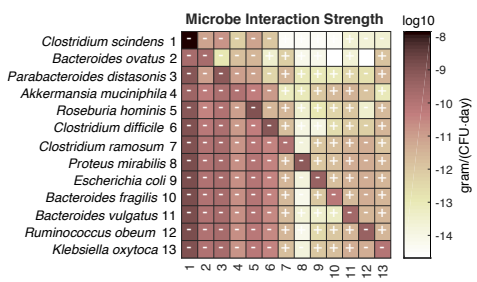


In vivo infection experiment

- Data from *C. difficile* challenge experiment performed with 5 germ free mice [Bucci et al. *Genome Biology* 2016]

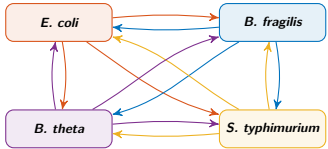






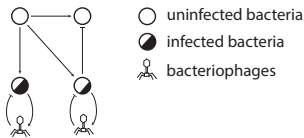
Three ongoing projects using the model

Marika Ziesack
Silver Lab, Harvard



- Microbes engineered to overproduces one amino acid
- Microbes engineered to need three amino acids
- Compare inference on WT and engineered strains to prove that engineering was performed.

Bryan Hsu
Silver Lab, Harvard



- Bacteriophages are "bacteria viruses"
- Using phages as control knobs to modulate microbiome
- Unlike antibiotics, phages are host specific and can be present in equilibrium with host

Massachusetts Host Microbiome Center

- Transplant complex bacteria into germ free mice (300+ bacteria)

Conclusions

1. Dynamical systems model for microbial dynamics based on what we term interaction modules (probabilistic clusters of latent variables with redundant interaction structure)
2. Bayesian formulation of the stochastic dynamical systems model that propagates measurement and latent state uncertainty throughout the model
3. Introduction of a temporally varying auxiliary variable technique to enable efficient inference by relaxing the hard non-negativity constraint on states

Time series metagenomics is a growing area for ML applications

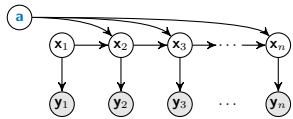
- Metagenomics is a "Big Data" problem, but number of time series samples is still sparse in comparison to the number of latent dynamical coefficients we need to learn.

Poster #2

Backup Slides

Simple example without the intermediate auxiliary variable

$$\mathbf{x}_{t+1,i} \mid \mathbf{x}_t, \mathbf{a} \sim \texttt{Normal}_{\geq 0}(\mathbf{a}_i{}^{\top} f(\mathbf{x}_t), \sigma^2_{\mathbf{x}_i})$$
$$\mathbf{y}_{t,i} \mid \mathbf{x}_{t,i} \sim \texttt{Normal}_{\geq 0}(\mathbf{x}_{t,i}, \sigma^2_{\mathbf{y}_i})$$
$$\mathbf{a}_i \sim \texttt{Normal}(0, \sigma^2_{\mathbf{a}_i})$$



Note the truncated distributions for **x** and **y**

Parameter inference Gibbs step: $\mathbf{a}^{(g+1)} \sim p_{\mathbf{a}|\mathbf{x}}(\cdot \mid \mathbf{x}^{(g)})$

$$p_{\mathbf{a}|\mathbf{x}} \propto p_{\mathbf{x}|\mathbf{a}} p_{\mathbf{a}}$$

$\texttt{Normal}_{\geq 0}(\mathbf{x}; \mu(\mathbf{a}, \mathbf{x}), \sigma^2)$ ↑ $p_{\mathbf{x}|\mathbf{a}}$; $p_{\mathbf{a}}$ ↓ $\texttt{Normal}(\mathbf{a}; 0, \sigma^2)$

$$= \frac{\mathbf{e}^{-\frac{1}{2\sigma^2}(\mathbf{x}-\mu(\mathbf{a},\mathbf{x}))^2}}{\sigma\sqrt{2\pi}\left(\Phi(\infty) - \Phi\left(-\frac{\mu(\mathbf{a},\mathbf{x})}{\sigma}\right)\right)} \frac{\mathbf{e}^{-\frac{1}{2\sigma^2}\mathbf{a}^2}}{\sigma\sqrt{2\pi}}$$

Sampling for other variables

- Filtering (sampling from posterior of **x**) is challenging
- Can not use collapsed Gibbs sampling for Dirichlet Process or Edge Selection

Negative Binomial

$$\mathbf{y}_{k,i} \mid \mathbf{q}_k \sim \texttt{NegBin}(\phi(\mathbf{q}_k, r_k), \epsilon(\mathbf{q}_k, a_0, a_1))$$

$$\phi(q_k, r_k) = r_k \frac{q_{k,i}}{\sum_i q_{k,i}} \tag{1}$$

$$\epsilon(q_k, a_0, a_1) = \frac{a_0}{q_{k,i}/\sum_i q_{k,i}} + a_1 \tag{2}$$

$$\texttt{NegBin}(y; \phi, \epsilon) = \frac{\Gamma(r+y)}{y!\,\Gamma(r)} \left(\frac{\phi}{r+\phi}\right)^y \left(\frac{r}{r+\phi}\right)^r$$

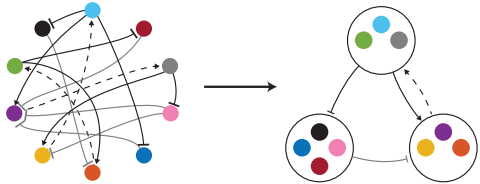
$$r = \frac{1}{\epsilon}$$

Priors

- We use conjugate priors on many variables (e.g., the variance terms $(\boldsymbol{\sigma}^2_{\mathbf{a}}, \boldsymbol{\sigma}^2_{\mathbf{b}}, \boldsymbol{\sigma}^2_{\mathbf{w}})$ have Inverse-Chi-squared priors).
- The module assignments, $\mathbf{c}$, are also updated by a standard Gibbs sampling approach for Dirichlet Processes.
- For the concentration parameter $\boldsymbol{\alpha}$, which has a Gamma prior on $\boldsymbol{\alpha}$, we use the sampling method described by Escobar and West 1995.

Backup: Alternative View of Bacteriotherapy

Interpretable inference for bacteriotherapy design



Predict microbial abundances (phenotype) in the presence of the bacteriotherapy