Robust and Scalable Models of Microbiome Dynamics for Bacteriotherapy Design

Travis E. Gibson[1] Georg K. Gerber[1,2]

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

[2] Health Sciences and Technology Division Harvard-MIT

December 9, 2017
NIPS 2017 Workshop on Machine Learning in Computational Biology

Outline

❶ Background on the Human Microbiome

❷ From Experimental Design to Bacteriotherapies

❸ Model of microbial dynamics

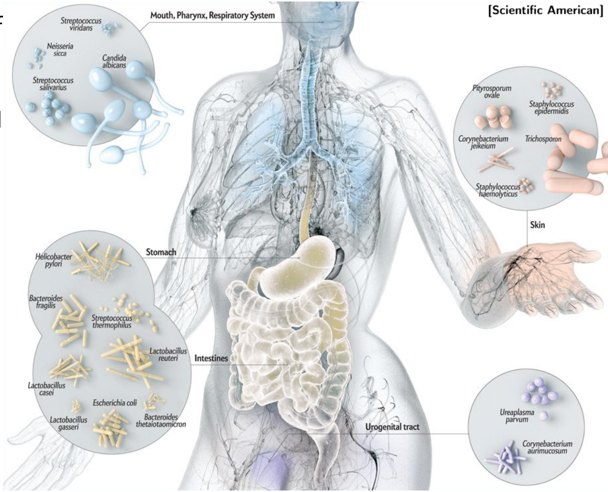
❹ Inference Model

❺ Applications

Gerber Lab is looking for Post-docs and PhD students

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. Large component of the immune system
5. Play a role in a variety of human diseases:
   - infections, arthritis, food allergy, cancer, inflammatory bowel disease, neurological diseases, and obesity/diabetes



[Scientific American]

Bacteriotherapy

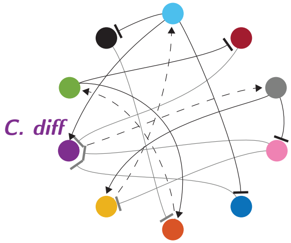
Bacteriotherapy: communities of bacteria administered to patients for specific therapeutic applications

- **“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

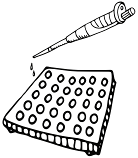
positive microbe A produces a small molecule (metabolite) that microbe B needs

negative two microbes competing for the same niche

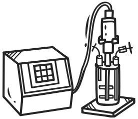
what if there were 300 bugs in the network?

Workflow in our lab

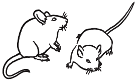
batch experiments



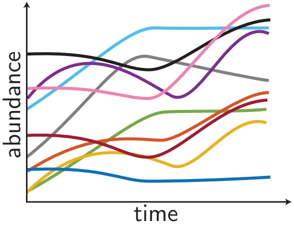
chemostat



animal experiments

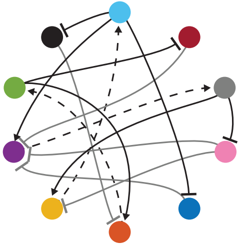


- 16S rRNA on MiSeq (reads) for relative abundances of species
- 16S rRNA qPCR (universal primers) for bacterial biomass



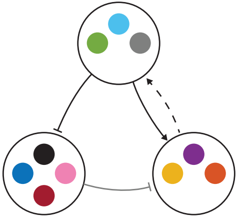
- measurements - irregular, sparse & noisy

Interaction Network



- 300 species
- 90,000 interactions

Interaction Modules



Microbial Dynamics

- 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, carrying capacity, interaction, stochastic disturbance

- Convert to discrete time

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

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

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

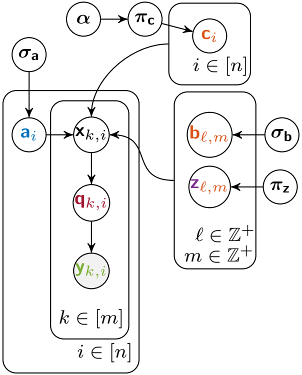
Dynamics

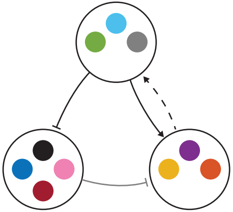
$$\mathbf{x}_{k+1,i} \mid \mathbf{x}_k, \mathbf{a}_i, \mathbf{b}, \mathbf{c}, \mathbf{z}, \boldsymbol{\sigma}_{\mathbf{w}} \sim \texttt{Normal}\left(\mathbf{x}_{k,i} + \mathbf{x}_{k,i}\left(\mathbf{a}_{i,1} + \mathbf{a}_{i,2}\mathbf{x}_{k,i} + \sum_{\mathbf{c}_j \neq \mathbf{c}_i} \mathbf{b}_{\mathbf{c}_i,\mathbf{c}_j}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

Constraint and Measurement Model

$$\mathbf{q}_{k,i} \mid \mathbf{x}_{k,i} \sim \texttt{Normal}(\mathbf{x}_{k,i}, \boldsymbol{\sigma}_{\mathbf{q}}^2)$$

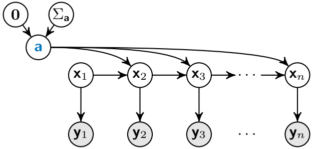
$$\mathbf{y}_{k,i} \mid \boldsymbol{\sigma}_{\mathbf{y}}, \mathbf{q}_{k,i} \sim f(\mathbf{q}_{k,i}) \quad f \in \{\texttt{Neg. Bin., Log Norm., ...}\}$$





Simple example without the intermediate auxiliary variable

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



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

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

$$p_{\mathbf{a}|\mathbf{x}} \propto p_{\mathbf{x}|\mathbf{a}} \overset{\substack{\mathtt{Normal}_{\geq 0}(\mathbf{x};\mu(\mathbf{a},\mathbf{x}),\sigma^2)\\ \uparrow}}{p_{\mathbf{x}|\mathbf{a}}} p_{\mathbf{a}} \underset{\substack{\downarrow\\ \mathtt{Normal}(\mathbf{a};0,\sigma^2)}}{p_{\mathbf{a}}}$$

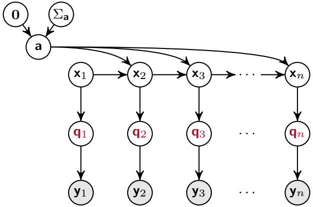
$$= \frac{\mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{x}-\mu(\mathbf{a},\mathbf{x}))^2}}{\sigma\sqrt{2\pi}\left(\Phi(\infty) - \Phi\left(-\frac{\mu(\mathbf{a},\mathbf{x})}{\sigma}\right)\right)} \frac{\mathrm{e}^{-\frac{1}{2\sigma^2}\mathbf{a}^2}}{\sigma\sqrt{2\pi}}$$

Sampling for other variables

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

Introducing an auxiliary variable

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



Prior on $\mathbf{q}$ is positive, relaxing the distribution on the dynamics for $\mathbf{x}$

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

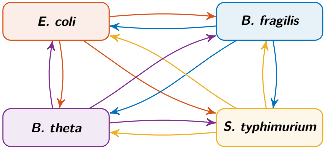
- Direct sampling from the posterior now possible (Bayesian Regression!)

Sampling for other variables

- Collapsed Gibbs sampling for Dirichlet Process and Edge Selection (integrate out $\mathbf{a}$)
- Filtering is still challenging but easier to design proposals than before (MH)

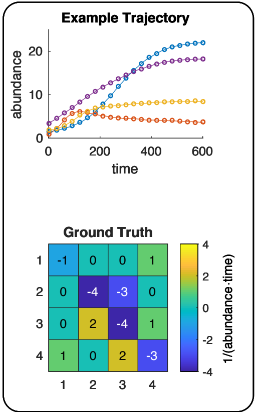
Synthetic consortia of small microbial community

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.

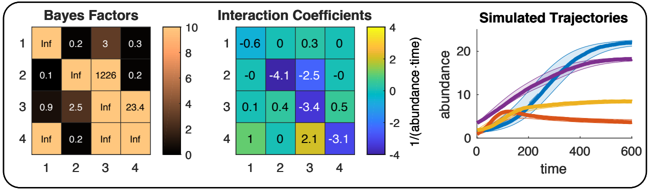
Synthetic Data



Ground Truth

	1	2	3	4
1	-1	0	0	1
2	0	-4	-3	0
3	0	2	-4	1
4	1	0	2	-3

Learning from 2 batch experiments



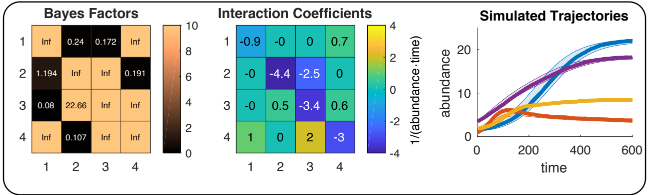
Bayes Factors

	1	2	3	4
1	Inf	0.2	3	0.3
2	0.1	Inf	1226	0.2
3	0.9	2.5	Inf	23.4
4	Inf	0.2	Inf	Inf

Interaction Coefficients

	1	2	3	4
1	-0.6	0	0.3	0
2	-0	-4.1	-2.5	-0
3	0.1	0.4	-3.4	0.5
4	1	0	2.1	-3.1

Learning from 4 batch experiments



Bayes Factors

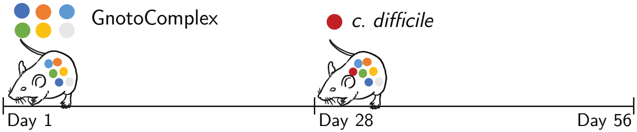
	1	2	3	4
1	Inf	0.24	0.172	Inf
2	1.194	Inf	Inf	0.191
3	0.08	22.66	Inf	Inf
4	Inf	0.107	Inf	Inf

Interaction Coefficients

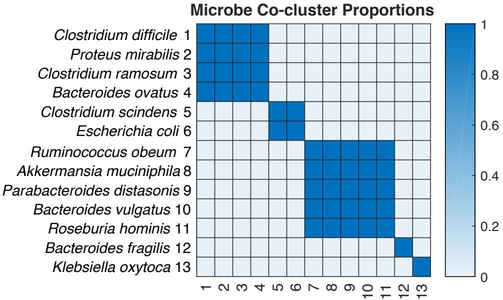
	1	2	3	4
1	-0.9	-0	0	0.7
2	-0	-4.4	-2.5	0
3	-0	0.5	-3.4	0.6
4	1	0	2	-3

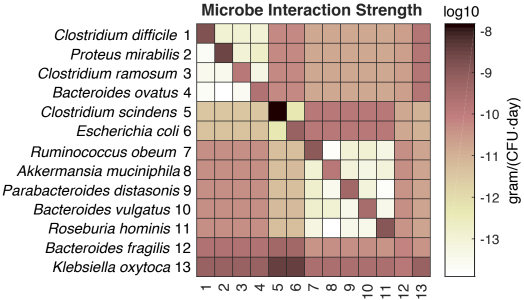
Animal experiments with *Clostridium difficile* infection

- Colonize mice with a defined complex of 12 bacteria (GnotoComplex), then challenge with *Clostridium difficile*



- 5 mice (26 fecal samples taken from each, 16s and universal qPCR)





Conclusions

We have presented

- Fully Bayesian inference model for microbial dynamics
- Interpretability features
    - Reducing the microbial interaction network complexity via extraction of modular features
    - Edge Selection so as to give us confidence as to what interactions are real

Future Directions

- Apply algorithm to mice that have been administered human fecal samples (complex flora 300+ species)
- Approximate Bayesian methods for dynamical systems analysis
- Modeling host dynamics (Layered latent dynamical processes)

Funding: DARPA, NIH
Acknowledge: Organizers, Gerber Lab, Bry Lab, Silver Lab

email: tgibson@mit.edu

Gerber Lab Plug

Gerber Lab is looking for post-docs and PhD students

Georg K. Gerber, MD, PhD, (ggerber@bwh.harvard.edu)

- Assistant Professor, Harvard Medical School
- Co-Director, Massachusetts Host-Microbiome Center
- Member of the Harvard-MIT Health Sciences & Technology Faculty
- Associate Pathologist, Center for Advanced Molecular Diagnostics Department of Pathology, Brigham and Women's Hospital