Lecture 9: Networks

SAMSI 2015: Program on Beyond Bioinformatics

Working group on Microbiome Community Dynamics and Complexity:

- Utility and dynamics of community state type (CST) definition for clinical applications;
- Reproducibility of sequencing;
- Vaginal microbiome dynamics and health;
- **Interaction inference in longitudinal data.**

Members of the Interaction inference subgroup

- Vanni Bucci, Dartmouth College
- Georg Gerber, Harvard University
- Christian Mueller, Simons Foundation
- Duncan Wadsworth, Rice University
- Min Wang, Iowa State University
- Chen Yanover, IBM

http://www.samsi.info/programs/2014-15-program-beyond-bioinformatics-statistical-and-mathematical-challenges-bioinformatic

2

Motivating example: Longitudinal study of microbiome in type 1 diabetes (T1D) conversion

Kostic et al., 2015, Cell Host & Microbe 17, 260–273

3

Inferring interactions between the bacteria

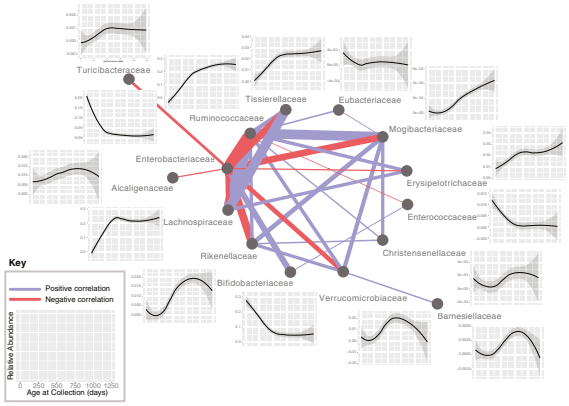


Figure 3. Temporal Dynamics of Microbial Taxonomies in Infant Gut Development

Family-level network diagram of the correlation between clades in their trajectories across time, excluding individuals with T1D. Positive correlations are in blue, negative correlations are in red, and the line thickness is proportional to the strength of the correlation (cumulative CCREPE Z statistic). The plots show the abundance of the indicated family as a smoothing spline across all healthy individuals with a 95% confidence interval (shaded region). See also Figure S2.

4

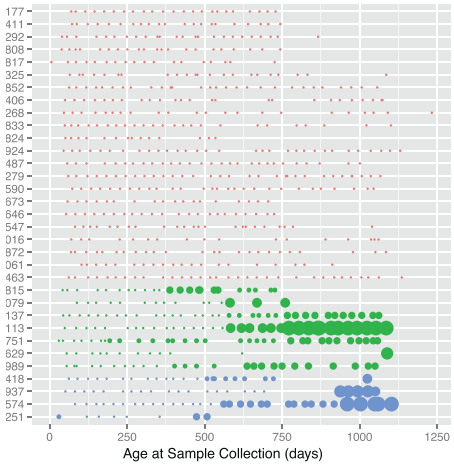
What is the significance of these interactions?

- How does conversion to T1D change the interactions between the bacteria?
- Are there unique early interaction patterns that can be related to T1D conversion?
- What do the inferred interaction tell us about the underlying generative processes governing T1D conversion?
- Are there any actionable points of intervention to prevent conversion?

5

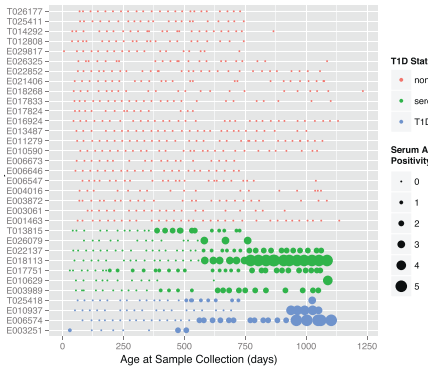
Properties of longitudinal microbiome datasets

- Asynchronous and uneven sampling
- Temporally sparse
- n << p (~1,000s of taxonomical groups/OTUs)
- Not Normally distributed
- Typically compositionally constrained data



6

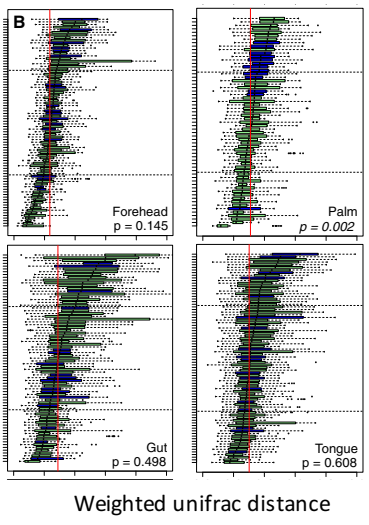
Kostic et al., 2015



- 33 subjects
- 11 seropositive for T1D
- 4 (out of 11) converted
- Sampling over 3 years
- ~20 samples

7

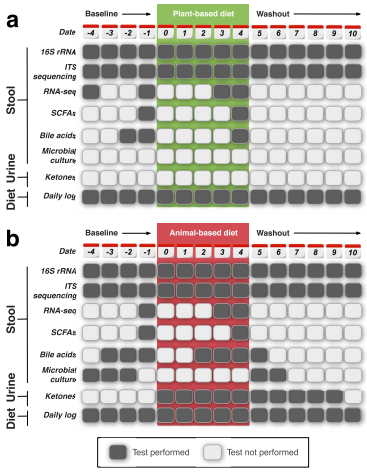
Flores et al., 2014



- 85 adults
- Weekly samples over 3 months (~36 per subject)
- Multiple body sites: forhead, gut, palm, tongue

8

David et al. 2014



- 11 Subjects
- Daily sampling 4 days before prescribed diet, 5 days during the diet, 6 days afterwards

9

Alternative analytic approaches

- Relevance/correlation networks (e.g. SPARCC)
- Graphical models (e.g. SPIEC-EASI)
- (Sparse) Vector autoregressive models (sVAR)
- Dynamical models (Generalized Lotka-Volterra)

10

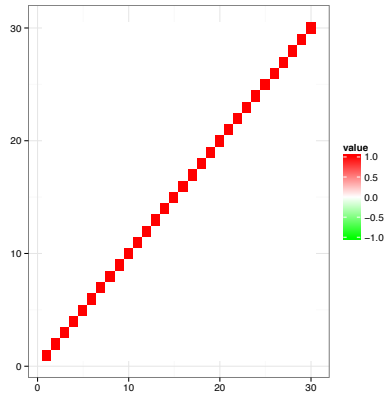
Relevance/Correlation networks

- Given variables X_1, …, X_p
- Calculate pairwise correlations, r_{ij} (or other pairwise association statistics)
- Infer an edge if r_{ij} is greater than some threshold
- E.g. in Kostic data before
- Generally, unreliable when n << p
- SPARCC method designed with sparsity assumptions (Friedman & Alm, 2012)

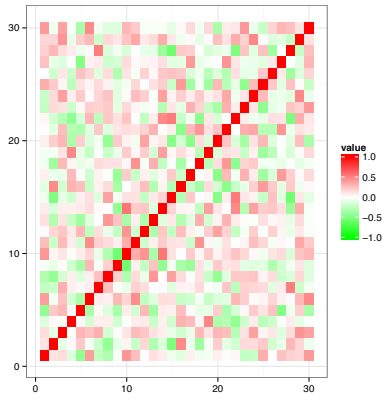
11

Sparse data may result in inference of spurious correlations/connections

Truth



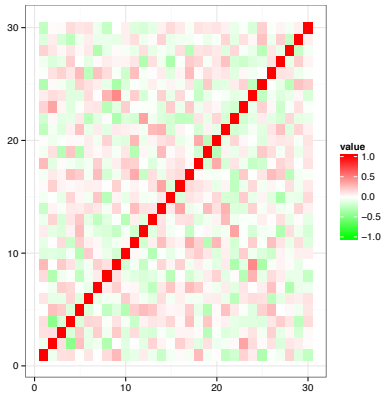
N=20



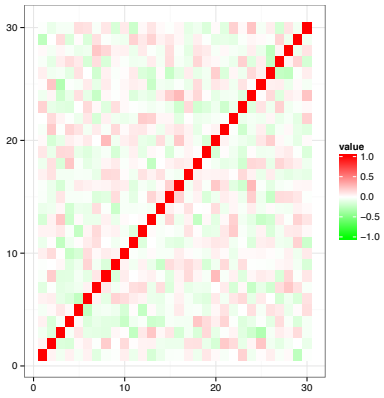
12

Empirical correlation matrices with no true dependencies

N=50



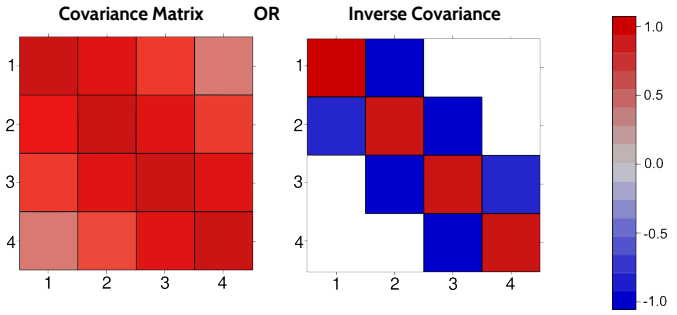
N=100



13

Conditional independence and sparsity

David MacKay's Gaussian Quiz. Assume a simple system of springs where you observe the position of the masses:



14

Graphical Models

- Infer inverse covariance C^{-1} matrix under regularizing sparsity assumptions
- C_{ij}^{-1} are non-zero iff X_i and X_j are conditionally dependent given all other variables
- SPIEC-EASI (arXiv: 1408.4158v3) optimize:

$$C^{-1} = \arg\min_{C^{-1} \in PD} -\text{logdet}(C^{-1}) + \text{tr}(C^{-1}S) + \lambda \|C^{-1}\|_1$$

15

Optimization algorithms for inverse covariance inference

- Neighborhood selection (Meinshausen and Buehlmann, 2006)
- Graphical LASSO (Yuan et al., 2007, Friedman et al. 2008, 2011)
- Alternating Linearization (Scheinberg et al., 2010)
- Quadratic Inverse Covariance (Hsieh et al, 2010)

16

Sparse Vector Autoregressive Models

- Let X_{ij} be abundance of microbe i at time j (row standardized)
- An order 1 VAR model is **X** = **Z b** + **e**, where
  - **b** is (p x p) coefficient matrix,
  - **Z** = (**X_1**, …, **X_{n-1}**)
  - **X** = (**X_2**, …, **X_n**).
- To infer **b** *l1* and *l2* regularization is applied.
- For asynchronously sampled data, smoothing and other pre-processing needs to be applied to make the data suitable for VAR models.

17

Dynamical models

- Models of the generalized Lotka-Volterra type can be specified to model the temporal changes in microbial abundance as a dynamical system
- $\frac{d}{dt}x_i(t) = \mu_i x_i(t) + x_i(t)\sum M_{ij}x_j(t) + x_i(t)\sum \varepsilon_{il}u_l(t)$
- Stein & Bucci et al. 2013 have developed regularized estimation procedure.
- MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses, Genome Biology, 2016.

18