Mixed Graphical Models for Microbiome and Metabolomic Data

Jing Ma

Public Health Sciences Division
Fred Hutch Cancer Research Center
jingma@fredhutch.org

August 1, 2018

Human Microbiome

- Communities of microbes that colonize all body surfaces.

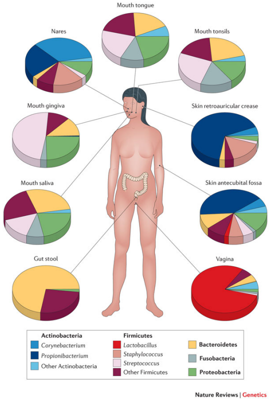


Fig: Compositional differences in human microbiome[1]

[1] Lasken and McLean, Nature Rev Genet, 2014

Human Microbiome

- Communities of microbes that colonize all body surfaces.
- Important in health and disease.

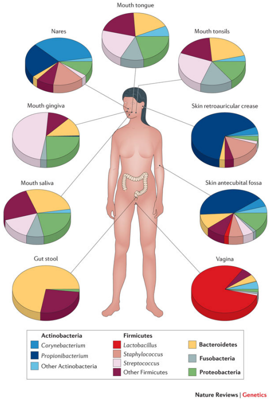


Fig: Compositional differences in human microbiome[1]

[1] Lasken and McLean, Nature Rev Genet, 2014

Human Microbiome

- Communities of microbes that colonize all body surfaces.
- Important in health and disease.
- More microbial cells than human cells.

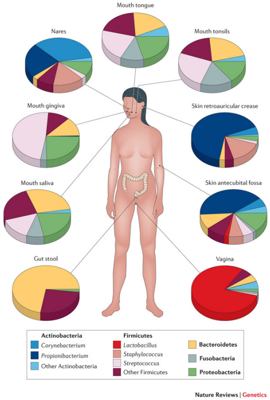


Fig: Compositional differences in human microbiome[1]

[1] Lasken and McLean, Nature Rev Genet, 2014

Human Microbiome

- Communities of microbes that colonize all body surfaces.
- Important in health and disease.
- More microbial cells than human cells.
- Who they are → What they are doing.

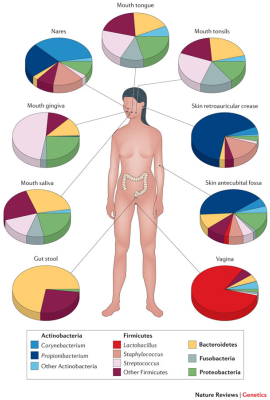
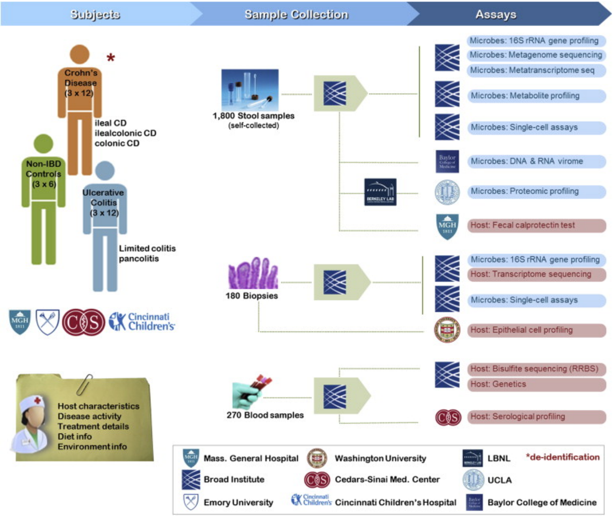


Fig: Compositional differences in human microbiome[1]

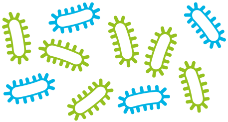
[1] Lasken and McLean, Nature Rev Genet, 2014

iHMP-IBD



iHMP Consortium. Cell Host Microbe. 2016

Metabolic Activity of the Microbiome



Gut bacteria

- Synthesize amino acids and vitamins
- Break down indigestible plant polysaccharides
- Produce metabolites involved in energy metabolism

Microbe - Metabolite Interactions

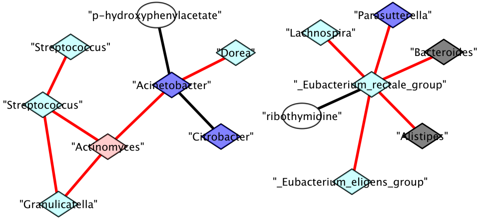
Problem of Interest

Can we use probabilistic graphical models to infer microbe-metabolite interactions from data?

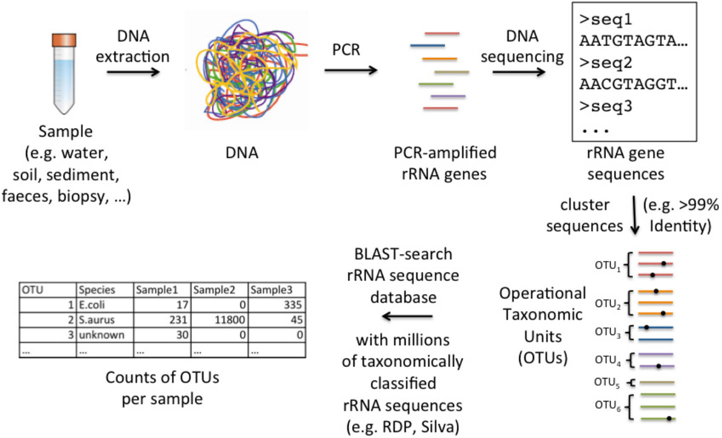
Microbe - Metabolite Interactions

Problem of Interest

Can we use probabilistic graphical models to infer microbe-metabolite interactions from data?

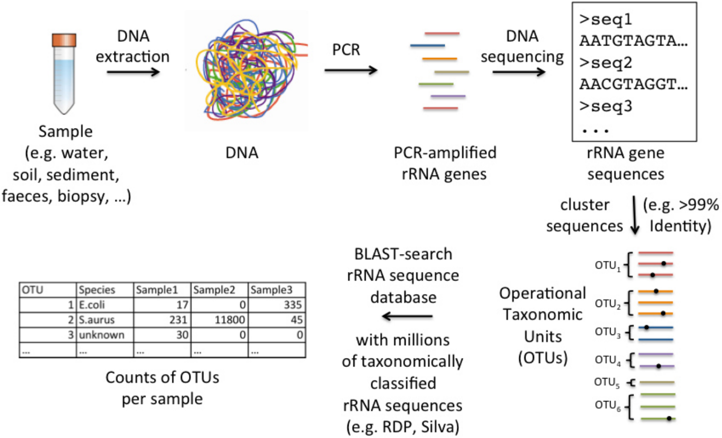


Microbiome Data



OTU	Species	Sample1	Sample2	Sample3
1	E.coli	17	0	335
2	S.aurus	231	11800	45
3	unknown	30	0	0
...	...	...	...	...

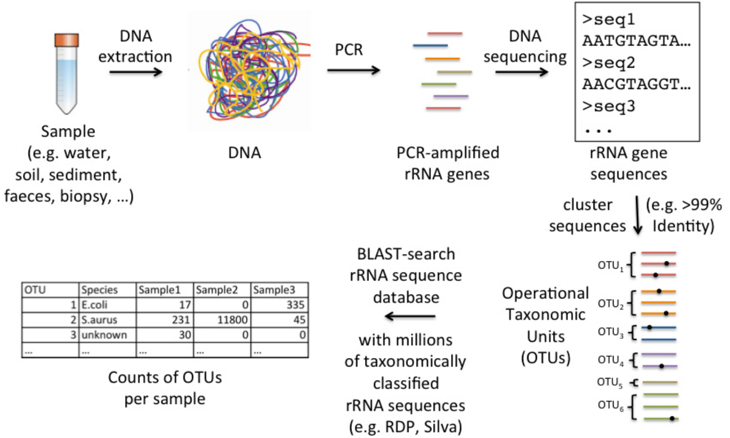
Microbiome Data



OTU	Species	Sample1	Sample2	Sample3
1	E.coli	17	0	335
2	S.aurus	231	11800	45
3	unknown	30	0	0
...	...	...	...	...

- OTU counts are noisy

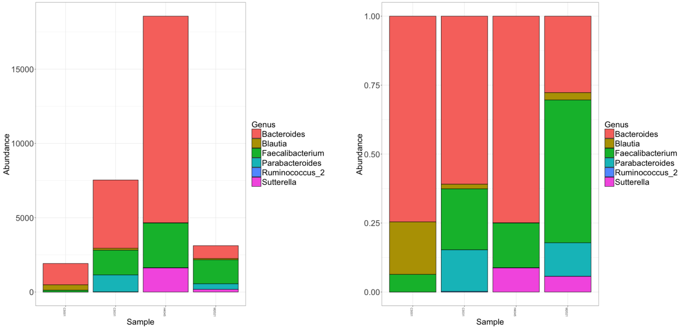
Microbiome Data



OTU	Species	Sample1	Sample2	Sample3
1	E.coli	17	0	335
2	S.aurus	231	11800	45
3	unknown	30	0	0
...	...	...	...	...

- OTU counts are noisy
- OTU matrix is sparse

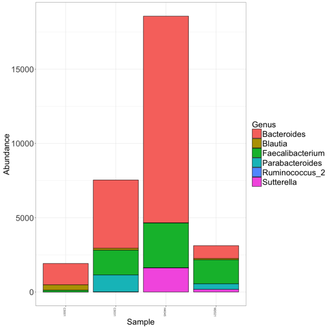
Microbiome Data

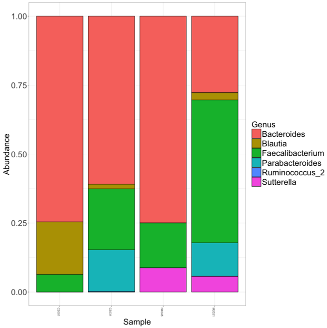


- Sequencing depth/library size varies.

[2] Kurtz et al. PLoS Comp Bio. 2015

Microbiome Data

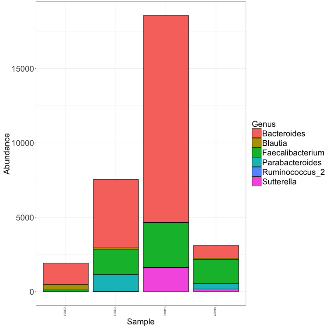


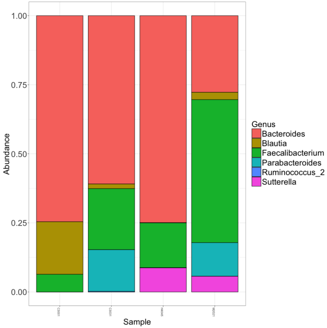


- Sequencing depth/library size varies.
- Existing networks are based on dissimilarity, correlation, graphical models[2].

[2] Kurtz et al. PLoS Comp Bio. 2015

Microbiome Data





- Sequencing depth/library size varies.
- Existing networks are based on dissimilarity, correlation, graphical models[2].
- CLR transformed data are not even close to Gaussian!

[2] Kurtz et al. PLoS Comp Bio. 2015

Our Framework

Key Idea: from compositional to ordinal data

0
0.1
0
0.12
0.4
0
0
0.28
0
0.1

Our Framework

Key Idea: from compositional to ordinal data

0	0
0.1	1
0	0
0.12	1
0.4	2
0	0
0	0
0.28	2
0	0
0.1	1

Probit Graphical Models

- Let $\boldsymbol{Y}^* \sim \mathcal{N}(0, \Sigma_{Y^*})$ be the latent variables and $Y_j^* \sim \mathcal{N}(0, 1)$.

W.l.o.g, assume $Y_j^* \sim \mathcal{N}(0, 1)$.

Probit Graphical Models

- Let $\boldsymbol{Y}^* \sim \mathcal{N}(0, \Sigma_{Y^*})$ be the latent variables and $Y_j^* \sim \mathcal{N}(0, 1)$.
- $\Sigma_{Y^*}^{-1}$ captures the conditional independence.

W.l.o.g, assume $Y_j^* \sim \mathcal{N}(0, 1)$.

Probit Graphical Models

- Let $\boldsymbol{Y}^* \sim \mathcal{N}(0, \Sigma_{Y^*})$ be the latent variables and $Y_j^* \sim \mathcal{N}(0,1)$.
- $\Sigma_{Y^*}^{-1}$ captures the conditional independence.
- Ordinal data $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ are discrete versions of $\boldsymbol{Y}^*$:

$$Y_j = \begin{cases} 0, & Y_j^* \in (-\infty, \theta_{1j}), \\ 1, & Y_j^* \in [\theta_{1j}, \theta_{2j}), \\ \vdots & \vdots \\ M-1, & Y_j^* \in [\theta_{M-1,j}, \infty). \end{cases}$$

W.l.o.g, assume $Y_j^* \sim \mathcal{N}(0,1)$.

Probit Graphical Models

- Let $\boldsymbol{Y}^* \sim \mathcal{N}(0, \Sigma_{Y^*})$ be the latent variables and $Y_j^* \sim \mathcal{N}(0,1)$.
- $\Sigma_{Y^*}^{-1}$ captures the conditional independence.
- Ordinal data $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ are discrete versions of $\boldsymbol{Y}^*$:

$$Y_j = \begin{cases} 0, & Y_j^* \in (-\infty, \theta_{1j}), \\ 1, & Y_j^* \in [\theta_{1j}, \theta_{2j}), \\ \vdots & \vdots \\ M-1, & Y_j^* \in [\theta_{M-1,j}, \infty). \end{cases}$$

- Θ and Σ_{Y^*} are unknown.

W.l.o.g, assume $Y_j^* \sim \mathcal{N}(0,1)$.

Why Ordinal?

FRED HUTCH
CURES START HERE™

- Discretization preserves key features of microbial interactions while mitigating noises

Why Ordinal?

- Discretization preserves key features of microbial interactions while mitigating noises
- Lead to a robust and more interpretable model
  - Conditional independence
  - $+$ partial correlation implies co-existence
  - $-$ partial correlation implies co-exclusiveness

Why Ordinal?

- Discretization preserves key features of microbial interactions while mitigating noises
- Lead to a robust and more interpretable model
  - Conditional independence
  - $+$ partial correlation implies co-existence
  - $-$ partial correlation implies co-exclusiveness
- Joint inference becomes easy

Mixed Graphical Models

- Mixed data $\underset{\text{ordinal}}{\boldsymbol{Y}}$ and $\underset{\text{con't}}{\boldsymbol{Z}}$
- $\boldsymbol{Y}$ is discrete version of $\boldsymbol{Y}^*$

Mixed Graphical Models

- Mixed data $\underset{\text{ordinal}}{\boldsymbol{Y}}$ and $\underset{\text{con't}}{\boldsymbol{Z}}$
- $\boldsymbol{Y}$ is discrete version of $\boldsymbol{Y}^*$
- The joint distribution of $(\boldsymbol{Y}^*, \boldsymbol{Z}) \sim \mathcal{N}(0, \Omega^{-1})$, where

$$\Omega^{-1} = \begin{pmatrix} \Sigma_{Y^*} & \Sigma_{Y^*Z} \\ \Sigma_{ZY^*} & \Sigma_Z \end{pmatrix}.$$

Mixed Graphical Models

FRED HUTCH CURES START HERE™

- Mixed data $\underset{\text{ordinal}}{\boldsymbol{Y}}$ and $\underset{\text{con't}}{\boldsymbol{Z}}$
- $\boldsymbol{Y}$ is discrete version of $\boldsymbol{Y}^*$
- The joint distribution of $(\boldsymbol{Y}^*, \boldsymbol{Z}) \sim \mathcal{N}(0, \Omega^{-1})$, where

$$\Omega^{-1} = \begin{pmatrix} \Sigma_{Y^*} & \Sigma_{Y^*Z} \\ \Sigma_{ZY^*} & \Sigma_Z \end{pmatrix}.$$

- Goal: infer Ω (and Θ) given i.i.d. $\{\boldsymbol{y}^{(i)}, \boldsymbol{z}^{(i)}\}$.

Estimation $\widehat{\Sigma}$

- Get $\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_{Y*} & \widehat{\Sigma}_{Y*Z} \\ \widehat{\Sigma}_{ZY*} & \widehat{\Sigma}_{Z} \end{pmatrix}$.
- Easy for $\widehat{\Sigma}_Z$!
- What about $\widehat{\Sigma}_{Y*}$ and $\widehat{\Sigma}_{Y*Z}$?

Estimation $\widehat{\Sigma}$

- Get $\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_{Y^*} & \widehat{\Sigma}_{Y^*Z} \\ \widehat{\Sigma}_{ZY^*} & \widehat{\Sigma}_{Z} \end{pmatrix}$.
- Easy for $\widehat{\Sigma}_Z$!
- What about $\widehat{\Sigma}_{Y^*}$ and $\widehat{\Sigma}_{Y^*Z}$?
- Estimate $\hat{\Theta}$

$$\hat{\theta}_{mj} = \Phi^{-1}(n^{-1}\sum_{i=1}^{n} \mathbf{1}(\boldsymbol{y}_j^{(i)} \leq m-1)) \quad m = 1, \ldots, M.$$

Estimation $\widehat{\Sigma}_{Y^*}$

- Estimate $\hat{\Sigma}_{jk}$

$$\hat{\Sigma}_{jk} = \underset{\sigma \in (-1,1)}{\arg\max}\, \ell_{jk}(\sigma; \hat{\Theta}),$$

where

$$\ell_{jk}(\sigma; \Theta) = \sum_{a=0}^{M} \sum_{b=0}^{M} \frac{n_{ab}}{n} \log \mathrm{P}(Y_j = a, Y_k = b; \Theta, \sigma)$$

and $n_{ab} = \sum_{i=1}^{n} \mathbf{1}(\boldsymbol{y}_j^{(i)} = a, \boldsymbol{y}_k^{(i)} = b)$.

Estimation $\widehat{\Sigma}_{Y*Z}$

- Estimate $\hat{\Sigma}_{j,p+k}$

$$\hat{\Sigma}_{j,p+k} = \underset{\sigma\in(-1,1)}{\arg\max}\, \ell_{j,p+k}(\sigma;\hat{\Theta}),$$

where

$$\ell_{j,p+k}(\sigma;\Theta) = \sum_{a=0}^{M} \frac{\sum_{i=1}^{n} \mathbf{1}(\boldsymbol{y}_j^{(i)} = a)}{n} \log \mathrm{P}(Y_j = a, \boldsymbol{z}_k^{(i)}; \Theta, \sigma).$$

Inference

- Apply graphical lasso

$$\widetilde{\Omega} = \operatorname*{arg\,min}_{\Omega \succ 0} \left\{ \mathrm{tr}(\widehat{\Sigma}\Omega) - \log\det(\Omega) + \lambda_n \|\Omega\|_{1,\mathrm{off}} \right\}$$

[3] Jankova and van de Geer. EJS. 2015

Inference

- Apply graphical lasso

$$\widetilde{\Omega} = \operatorname*{arg\,min}_{\Omega \succ 0} \left\{ \mathrm{tr}(\widehat{\Sigma}\Omega) - \log\det(\Omega) + \lambda_n \|\Omega\|_{1,\mathrm{off}} \right\}$$

- Debias[3]

$$\widehat{\Omega} = 2\widetilde{\Omega} - \widetilde{\Omega}\widehat{\Sigma}\widetilde{\Omega}$$

[3] Jankova and van de Geer. EJS. 2015

Denote $p' = \max\{p + q, n\}$ and $s_0 = \#\{\Omega_{jk} \neq 0 : 1 \leq j < k \leq p + q\}$.

Theory

Denote $p' = \max\{p+q, n\}$ and $s_0 = \#\{\Omega_{jk} \neq 0 : 1 \leq j < k \leq p+q\}$.

Theorem

Under some regularity conditions on $\Omega^*, \mathrm{P}(Y_j = a, Y_k = b; \Theta^*, \sigma)$ *and* $\mathrm{P}(Y_j = a, \boldsymbol{z}_k; \Theta^*, \sigma)$, *for* $n \gtrsim s_0^2 \log p'$ *and* $\lambda_n = O(\sqrt{\log p'/n})$, *we have w.h.p*

$$\max_{j,k} |\widehat{\Sigma}_{jk} - \Sigma^*_{jk}| \leq \sqrt{\frac{\log p'}{n}}.$$

Theory

Denote $p' = \max\{p+q, n\}$ and $s_0 = \#\{\Omega_{jk} \neq 0 : 1 \leq j < k \leq p+q\}$.

Theorem

Under some regularity conditions on $\Omega^*, \mathrm{P}(Y_j = a, Y_k = b; \Theta^*, \sigma)$ *and* $\mathrm{P}(Y_j = a, \boldsymbol{z}_k; \Theta^*, \sigma)$, *for* $n \gtrsim s_0^2 \log p'$ *and* $\lambda_n = O(\sqrt{\log p'/n})$, *we have w.h.p*

$$\max_{j,k} |\widehat{\Sigma}_{jk} - \Sigma^*_{jk}| \leq \sqrt{\frac{\log p'}{n}}.$$

Intuition:

- $\widehat{\Sigma}_{jk}$: empirical loss function $\ell_{jk}(\cdot)$ is non-convex.
- Assumptions ensure a one-to-one correspondence between critical points of the empirical loss and the population loss.

Denote $s_{jk}^2 = \Omega_{jj}^* \Omega_{kk}^* + \Omega_{jk}^{*2}$.

Corollary

Under an additional irrepresentable condition on $\Sigma^* \otimes \Sigma^*$

$$\sqrt{n}(\widehat{\Omega}_{jk} - \Omega_{jk}^*)/s_{jk} = W_{jk}^n + o_p(1),$$

where W_{jk}^n converges weakly to $\mathcal{N}(0, 1)$.

FRED HUTCH CURES START HERE™

Denote $s_{jk}^2 = \Omega_{jj}^* \Omega_{kk}^* + \Omega_{jk}^{*2}$.

Corollary

Under an additional irrepresentable condition on $\Sigma^* \otimes \Sigma^*$

$$\sqrt{n}(\widehat{\Omega}_{jk} - \Omega_{jk}^*)/s_{jk} = W_{jk}^n + o_p(1),$$

where W_{jk}^n converges weakly to $\mathcal{N}(0,1)$.

Structural Recovery:

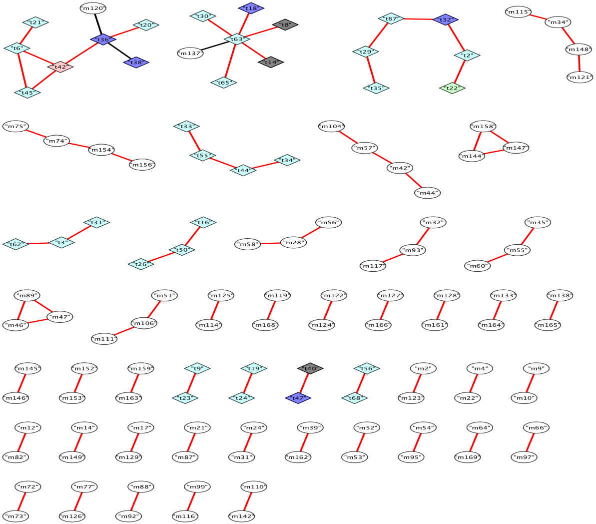
- For each pair $1 \leq j < k \leq p+q$, test $H_{0,j,k} : \Omega_{jk} = 0$
- Correct for multiple testing via BH

Multi-Omics Analysis of IBD

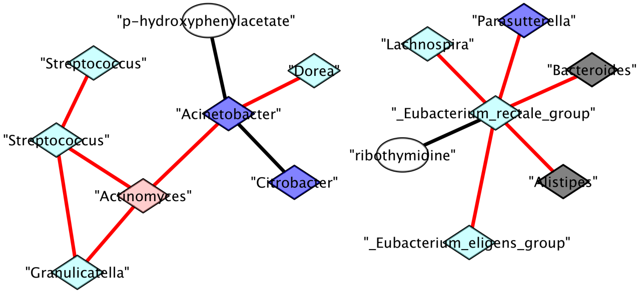
- Number of subjects $n = 81$
- 982 OTUs $\rightarrow$ $p = 68$ after removing sparse ones
- Discretization: use 0 and 67% quantile (M=3)
- 304 metabolites $\rightarrow$ $q = 169$ after removing those with small correlations
- Visualize the top 81 most significant edges

Results (Colored Nodes: Taxa)

FRED HUTCH
CURES START HERE™



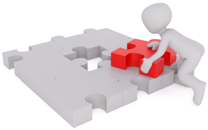
Results (Colored Nodes: Taxa)



- Edges colored in red represent positive partial correlations.
- Two nodes named Streptococcus have distinct OTU IDs.
- *Acinetobacter sp.* are capable of converting p-hydroxyphenylacetate into biochemical metabolites necessary for their growth[4].

[4] Thotsaporna et al. J Mol Catal B Enzym. 2016

Summary



- A framework for joint analysis of microbiome and metabolomic data using mixed graphical models
- An inferential procedure for uncertainty quantification of each interaction

Thank You

Link to Latent Variable Graphical Model[5]

$$\begin{pmatrix} \Sigma_{Y^*} & \Sigma_{Y^*Z} \\ \Sigma_{ZY^*} & \Sigma_Z \end{pmatrix}^{-1} = \begin{pmatrix} \Omega_{Y^*} & \Omega_{Y^*Z} \\ \Omega_{ZY^*} & \Omega_Z \end{pmatrix}$$

- Let Z be observed and Y^* be hidden variables.
- Schur complement

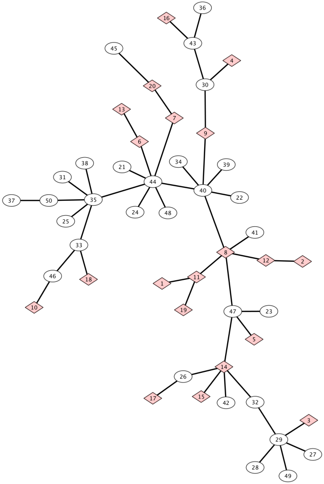
$$\Sigma_Z{}^{-1} = \underbrace{\Omega_Z}_{\text{sparse}} - \underbrace{\Omega_{ZY^*}(\Omega_{Y^*})^{-1}\Omega_{Y^*Z}}_{\text{low-rank}}$$

- We assume knowledge of Y^* in the form of ordinal variables whereas Chandrasekaran et al. (2012) assumes no knowledge of Y^*.

[5] Chandrasekaran et al. Ann. Statist. 2012

Simulation

- A scale-free network
- Colored nodes are ordinal
- Generate $(\mathbb{Y}^*, \mathbb{Z})$ from Ω
- $M = 3$
- $\theta_{mj} \in \{\pm 0.5, \pm 0.8\}$
- Generate $\mathbb{Y}$ from $\mathbb{Y}^*$ and Θ
- Estimate Ω from $(\mathbb{Y}, \mathbb{Z})$



$n = 100, p = 40, q = 60$

BH correction with $\alpha = 0.25$