Discovery of metabolic cooperation in complex microbial communities

Machine Learning (CSE 546) Project

Janet Matsen (Chemical Engineering) and Serena Liu (Genome Sciences)

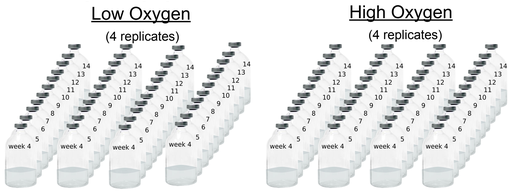
Background

How do bacteria cooperate to remediate methane in nature? What contributions do microbes that cannot metabolise methane add? To investigate metabolic interactions in diverse populations and the genes that drive these interactions, we analyzed RNA-sequencing (RNA-seq) data collected from cultures of Lake Washington sediment.

Sediment was collected from Lake Washington using a sediment corer (Fig. 1a). The samples were then cultured in the lab with methane, and either high or low oxygen, for 14 weeks. Metagenomes were sampled for weeks 4-14. Relative microbe abundances and gene expression levels were inferred from this 5 terabyte sequencing data set.

Figure 1. Experimental setup & equipment

(a) Sediment corer used to gather the Lake Washington samples



(b) Diagram of the experimental setup: 8 independent sediment samples were collected from Lake Washington. 4 were cultured in low oxygen conditions, and 4 in high oxygen conditions. Samples were serially transferred weekly, and metagenome RNA-seq data was collected for weeks 4-14 for each sample.

We focus on two groups of organisms in these samples: methanotrophs (can metabolize methane) and methylotrophs (cannot metabolize methane). Both the fraction assigned to each category and the composition of organisms within each category varies across samples. We aim to determine which genes drive differences in microbial population compositions.

Data Challenges

Data are not independent and identically distributed (iid)

- Time series data → samples in each series are sequentially dependent
- Differing growth conditions → samples are not identically distributed

Small N, big d (It's biology!)

- 83 samples & 212,710 features
- Affects which questions we asked & model selection
- Caution required to prevent over-fitting of models.

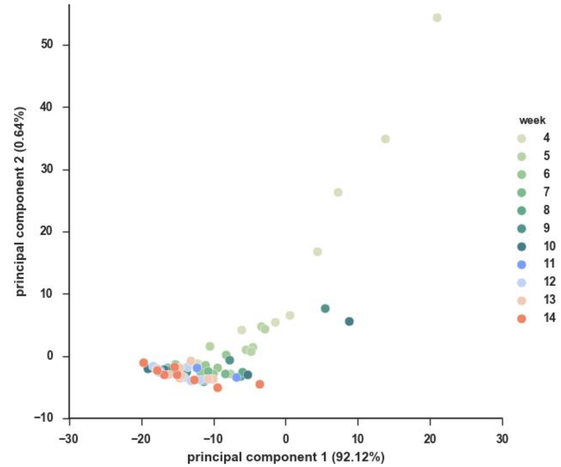
Methods

Data preprocessing

- Aggregated gene expression data for individual species based on gene name within methanotrophs and methylotrophs:
  - 44,210 & 149,137 features → 5,055 & 12,030 features
- Removed genes with zero variance (for CCA)
  - 5,055 & 12,030 features → 4,840 & 10,123 features
- Removed "hypothetical" genes and genes of unknown function
  - 4,840 & 10,123 features → 4,563 & 9,324 features
- Normalized expression data to have mean 0 and unit variance

PCA to assess trends in series and time

- Individual replicates are generally tightly correlated with each other
- Observe time dependent effects for early vs. late samples



Sparse CCA

Canonical correlation analysis (CCA): approach for multivariate analysis of two datasets

- Useful for identifying relationships among complex data

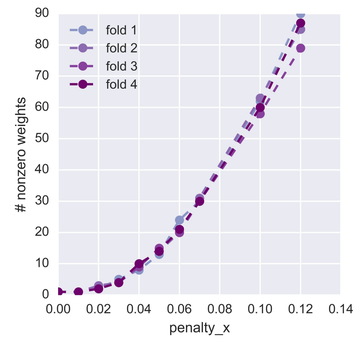
Standard CCA is not well defined when $N < \min(d_1, d_2)$

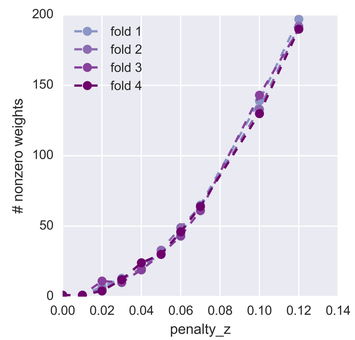
- We have lots of features → sparse CCA (L1-penalized variant of CCA)
- Used the R package PMA[1]

Tune hyperparameters such that we maximize validation set correlation in the projected space

Results

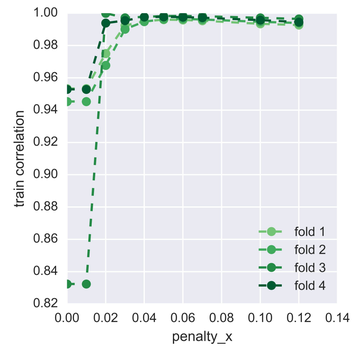
Hyperparameter tuning

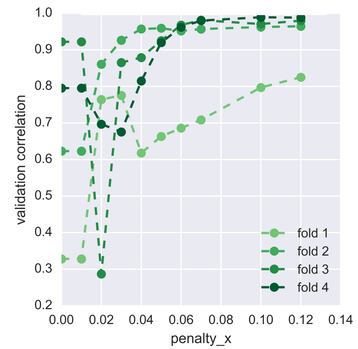




- Lower penalty parameters correspond to more stringent regularization → fewer nonzero weights
- X: methanotroph genes
- Z: methylotroph genes

Cross-Validation





References

1. Witten, D. M., Tibshirani, R., and Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* (2009).

Learn more at https://github.com/JanetMatsen/ML_meta-omics

Seattle, WA, December 2016