SIMONS FOUNDATION

Interpretable Feature Selection in Metagenomic Data

Advancing Genomic Biology Through Novel Method Development

Cara Magnabosco
June 5, 2017

What identifies microbial communities?

Basic setup:
Given a count matrix, can we identify the features that best predict the response?

Features (e.g. OTUs, KEGG categories, proteins, etc)

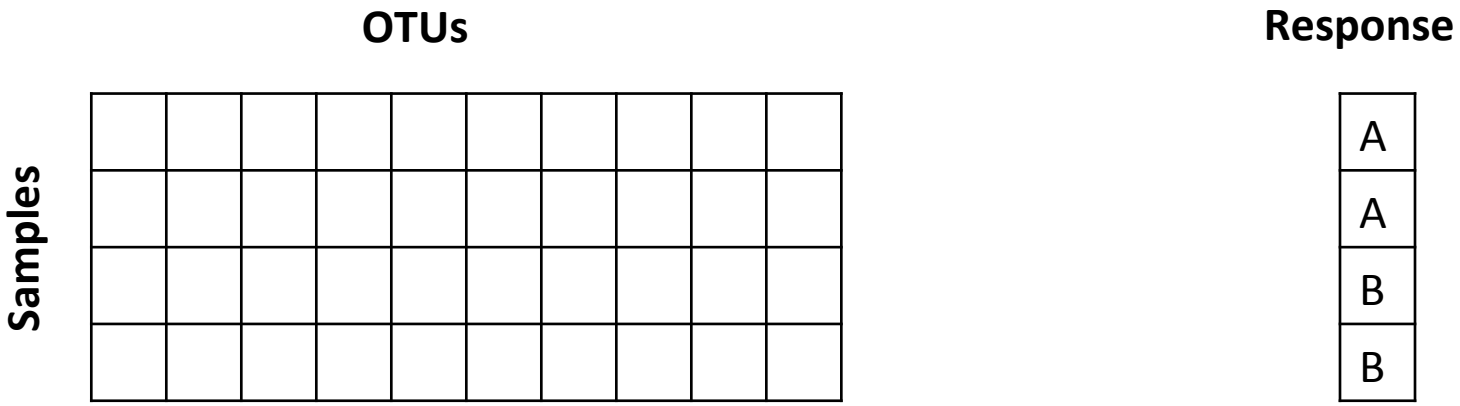
Samples

Response

A
A
B
B

SIMONS FOUNDATION

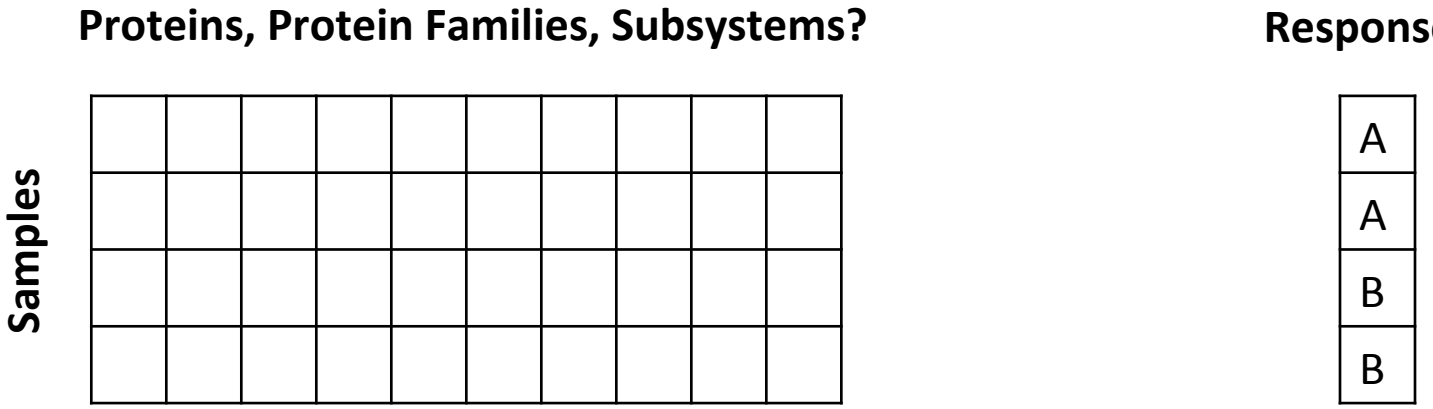
1a) This is “solved” for 16S datasets



e.g. A popular method:
Linear discriminant analysis effect size (LEfSe; Segata et al. 2011)

cmagnabosco@flatironinstitute.org

1b) How do we build the count matrix for metagenomes?



- Which features do we use?

- How to build the count matrix efficiently?

- How to deal with missing data and the influence of community composition?

cmagnabosco@flatironinstitute.org

2) Too broad or too specific, which features to use?

KEGG Hierarchy, Protein Families, Proteins

Functional Category ⟷ Phylogenetic Info.

Choice depends how different/similar the samples you are comparing

SIMONS FOUNDATION

cmagnabosco@flatironinstitute.org

3) Hundreds of samples, billions of reads. How to construct the count matrix efficiently?

KEGG Hierarchy, Protein Families, Proteins

Functional Category ⟷ Phylogenetic Info.

Read mapping/Assembly/Binning/Clustering

HMMER/BLAST ⟷ Short Read Placement

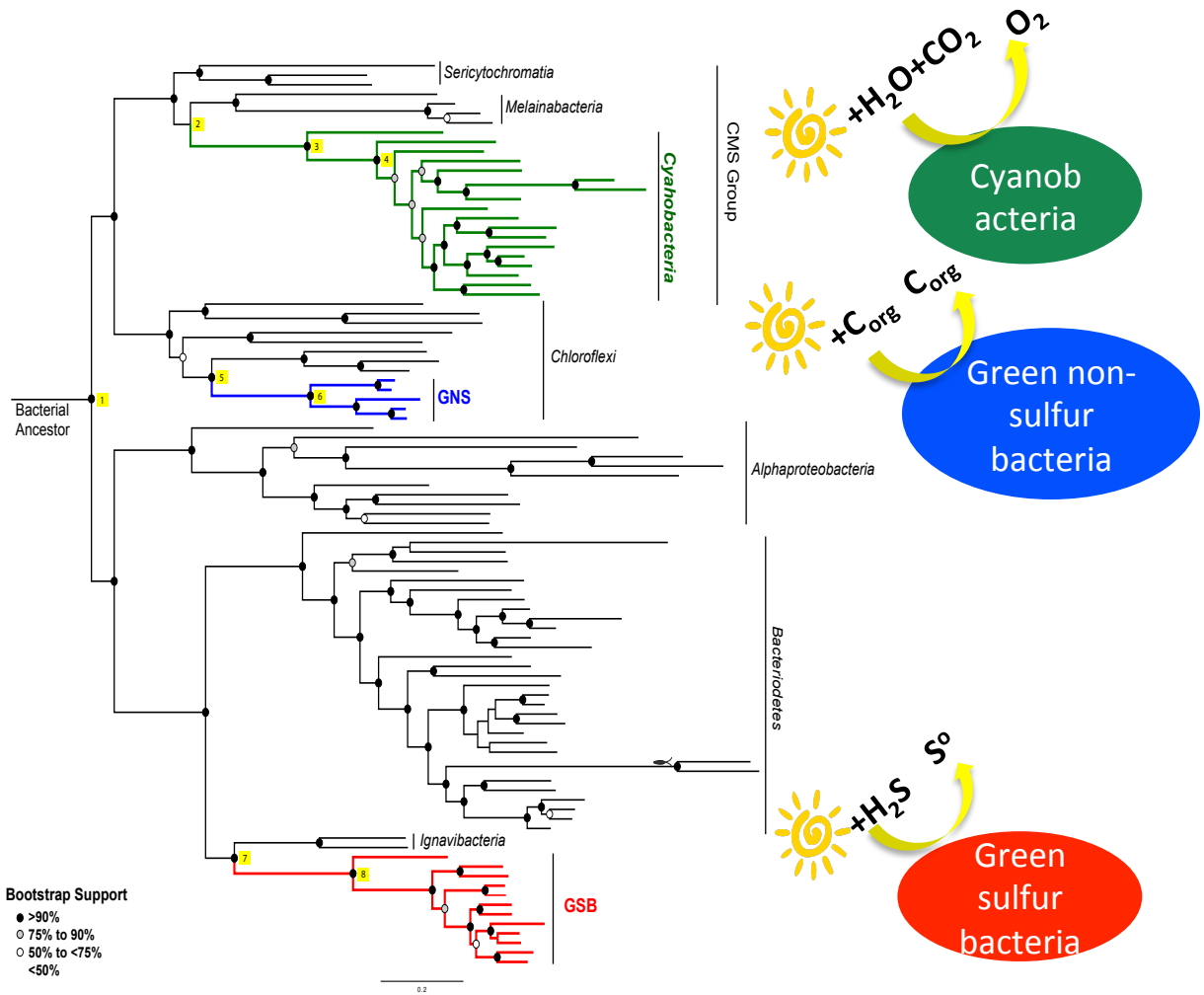
1. Can I easily add more samples?
2. How does the method handle sequences that aren't represented in the database?
3. What about missing data? Copy number issues?

SIMONS FOUNDATION

cmagnabosco@flatironinstitute.org

4) Getting information from metagenomes not in the 16S

Did I find something new, or did I just rediscover that Cyanobacteria have photosystems?



16S can tell us a lot about which individuals are present or enriched given a condition but the next step in the problem is understanding why.

SIMONS FOUNDATION

cmagnabosco@flatironinstitute.org