Prediction strength

Back to Table of Contents

All of the code in this page is meant to be run in `R` unless otherwise specified.

Loading a genus table and the metadata into R

Before loading data into R, this QIIME command must be run on the command line to collapse OTU counts into genus (L6) and phylum (L2) count tables:

```
# (run on command line)
summarize_taxa.py -i otu_table.biom -L 6

# convert to JSON BIOM format to load into R using R biom package:
biom convert -i otu_table_L6.biom -o otu_table_L6_json.biom --to-json
```

Inside `R`, Install biom, vegan, and cluster packages **if not installed**.

```
install.packages(c('biom','vegan','cluster'),repo='http://cran.wustl.edu')
```

Load packages

```
library('biom')
library('vegan')
library('cluster')
```

Load data

```
# load biom file
genus.biom <- read_biom('otu_table_L6_json.biom')

# Extract data matrix (genus counts) from biom table
genus <- as.matrix(biom_data(genus.biom))

# transpose so that rows are samples and columns are genera
genus <- t(genus)

# load mapping file
map <- read.table('map.txt', sep='\t', comment='', head=T, row.names=1)
```

It is extremely important to ensure that your genus table and metadata table sample IDs are lined up correctly.

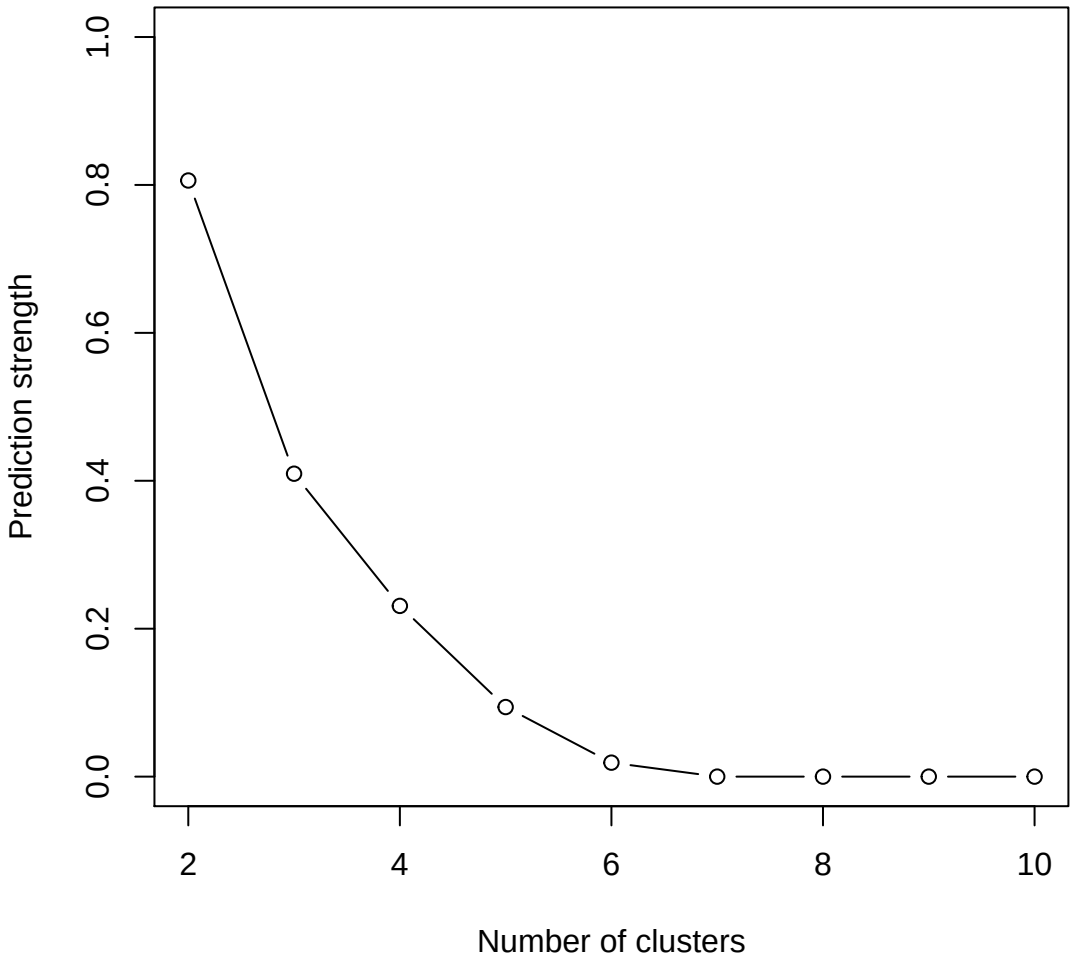
```
# find the overlapping samples
common.ids <- intersect(rownames(map), rownames(genus))

# get just the overlapping samples
genus <- genus[common.ids,]
map <- map[common.ids,]
```

Calculate prediction strength

```
# Source the prediction strength code.
source('../../src/prediction.strength.r')

# Calculate Bray-Curtis distances
bc <- vegdist(genus)

# Run prediction strength analysis on bray-curtis table
# Use 100 random splits
ps <- prediction.strength(bc,M=100)

# print the prediction strength values
cat(paste("K =",2:10," Prediction Strength =", round(ps$ps.mean,3), "+/-", round(ps$ps.sd,3), sep=' ',c
```

```
## K = 2  Prediction Strength = 0.806 +/- 0.124
## K = 3  Prediction Strength = 0.41 +/- 0.112
## K = 4  Prediction Strength = 0.231 +/- 0.128
## K = 5  Prediction Strength = 0.094 +/- 0.101
## K = 6  Prediction Strength = 0.019 +/- 0.06
## K = 7  Prediction Strength = 0 +/- 0
## K = 8  Prediction Strength = 0 +/- 0
## K = 9  Prediction Strength = 0 +/- 0
## K = 10  Prediction Strength = 0 +/- 0
```

```
# plot prediction strength
plot(2:10,ps$ps.mean,type='b',ylim=c(0,1),xlab='Number of clusters', ylab='Prediction strength')
```



Here we see that USA vs. non-USA does not have strong discrete cluster structure at the genus level according to prediction strength.

Make a PCoA plot colored by cluster, but showing COUNTRY by shape of the points. Note that USA is clustered almost perfectly.

```
# Run partitioning around medoids (PAM) clustering with 2 clusters
p <- pam(bc,2)

# PCoA coordinates of Bray-Curtis distances
pc <- cmdscale(bc,2)

# plot PCoA colored by cluster, with countries shown by shape.
plot(pc[,1], pc[,2], col=c('blue','red')[p$clustering], pch=c(16,17,18)[map$COUNTRY])
legend('topleft',legend=c('USA','Malawi','Venezuela'),pch=c(17,16,18))

# Plot the cluster labels at the centroids
cluster.centroids.x <- sapply(split(pc[,1],p$clustering),mean)
```

```
cluster.centroids.y <- sapply(split(pc[,2],p$clustering),mean)
text(cluster.centroids.x, cluster.centroids.y, c('Cluster 1', 'Cluster 2'),col=c('blue','red'))
```