Week 15, Lecture 29

István Albert
Bioinformatics Consulting Center
Penn State
Metagenomics: data analysis

• It is a new application domain

• This has benefits but also frustration

• Historically: grandiose statements for nascent technologies are usually exaggerated
Two big packages

• **QIIME** – Quantitative Insights Into Microbial Ecology (Nature Methods, 2010)
  – it is a Python based “glue” that connects a series of external tools
  – Has a very long list of programs that need to be installed: blast, rdp, CD-hit, uclust, fasttree etc

• **mothur** (Appl. Environ Microbiology, 2009) – by Patrick Schloss – a single binary file that contains **all the functionality**
Lot of authors, published in Nature Methods, 2010
Has a Google Group forum that seem very active
Welcome to the website for the mothur project, initiated by Dr. Patrick Schloss and his software development team in the Department of Microbiology & Immunology at The University of Michigan. This project seeks to develop a single piece of open-source, expandable software to fill the bioinformatics needs of the microbial ecology community. In February 2009 we released the first version of mothur, which had accelerated versions of the popular DOTUR and SONS programs. Since then we have added the functionality of a number of other popular tools including s-libshuff, TreeClimber (i.e. the parsimony test), UniFrac, distance calculation, visualization tools, a NAST-based aligner, and many other features. If you would like to contribute code to the project feel free to download the source code and make your own improvements. Alternatively, if you have an idea or a need, but lack the programming expertise, let us know and we'll add it to the queue of features we would like to add. Our current goal is to release a new iteration of the project monthly.
mothur is a unique and amazing software

• There is no other bioinformatics software like it → a design reminiscent of **matlab** or **mathematica**

Written by a tiny team, yet implements very advanced methods and techniques!

• You can finish the entire analysis/paper by using only mothur (plot with R)

• Excellent documentation (though some aspects refer to previous mothur releases so may be out of sync)
Mothur manual

The goal of mothur is to have a single resource to analyze molecular data that is used by microbial ecologists. Many of these tools are available elsewhere as individual programs and as scripts, which tend to be slow or as web utilities, which limit your ability to analyze your data. mothur offers the ability to go from raw sequences to the generation of visualization tools to describe α and β diversity. Examples of each command are provided within their specific pages, but several users have provided several analysis examples, which use these commands. An exhaustive list of the commands found in mothur is available within the commands category index. If you have any questions, complaints, or praise, please do not hesitate to email Pat or to use the various discussion tabs:

- General operations
- Sequence processing
- OTU-based approaches
- Hypothesis testing approaches
- Frequently asked questions
- Phylotype analysis

OTU based approaches

• Characterize similarity and frequency distribution of the sequences to evaluate ecological features:
  
  – Richness
  – Diversity
  – Similarity

Depends on our ability to compare sequences!
NAST – Nearest alignment space termination algorithm

• DeStantis et al, Nucleic Acid Research (2006)

Creates MSA (multiple sequence alignments)

1. The template sequences are pre-aligned relative to one another.

2. Then query (candidate) sequences will also be aligned to the MSA and therefore comparable relative to one another.
Example of NAST compression of a BLAST pairwise alignment using a 38 character aligned template.

DeSantis T Z et al. Nucl. Acids Res. 2006;34:W394-W399
Distances in a MSA

• We can compute the number of mismatches + gaps between any two sequences

• Once we have a distance we can cluster with this metric

• Sequences that cluster at a certain level form a operational taxonomical unit
Metagenomics “calculators”

Analogous to the game “mastermind”

One tries to deduce the truth with simple statements/guesses that reduce the possibilities after each step.

Example:

• Are the abundances the same
  • Yes $\rightarrow$ no difference between samples
  • No $\rightarrow$ is the dominance of the most common species the same?
    1. Yes $\rightarrow$ so the differences all come from the less common species
    2. No $\rightarrow$ is the dominance different because there are fewer species, more species, or same but with different abundances
Calculators in mothur

The following are a list of calculators currently available in mothur for OTU-based analyses. If there is a calculator that you think should be added, please consider either contributing your code or asking us to add it to our list of features for future versions of mothur. Following the links will take you to pages describing the formulae and example calculations using the Amazonian dataset.

Contents [hide]

1. Community richness
2. Community evenness
3. Community diversity
4. Estimates of number of additional OTUs observed with extra sampling
5. Statistical distributions
6. Shared community richness
7. Similarity in community membership
8. Similarity in community structure
9. Utility calculators

Community richness

- sobs - the observed richness
- chao - the Chao1 estimator
- ace - the ACE estimator
- jack - the jackknife estimator
- bootstrap - the bootstrap estimator
Simpson’s paradox

Most famous example: berkley gender bias lawsuit

<table>
<thead>
<tr>
<th>Department</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>Admitted</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>62%</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63%</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37%</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33%</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28%</td>
</tr>
<tr>
<td>F</td>
<td>272</td>
<td>6%</td>
</tr>
</tbody>
</table>

Total admission rates seem to statistically favor men.

Admission rates per largest departments seem to statistically favor women.

Difficult to recognize this class of problems – yet we face them all the time.
In meta-genomics where we have to constantly compare across groups, memberships and abundances.
Proper experimental design

- Metagenomics studies critically depend on controls and experimental designs – moreso than any other study.

- The parameter space is huge:
  - contaminants, chimeric reads, sequencing errors will all look like a new bacterial species
  - the data is sparse and noisy – lots of dead ends

- Search for: SOP on the mothur webpage (SOP – standard operating procedure)

1. Advocates dedicating two barcodes to two different controls
   a) mock (prebuilt community) and
   b) known realistic sample that you always re-sequence for every experiment
mothur in action

```r
# usage:
# mothur script-name

# make mothur append into the same log
set.logfile(name="work.log", append=T)

# get a summary of sequences
summary.seqs(fasta="x.fasta")

# splitting by barcodes (also here we would filtering by quality)
trim.seqs(fasta="x.fasta", oligos="x.oligos")

# get the unique sequences
unique.seqs(fasta="x.fasta")

# get a summary of the unique sequences
summary.seqs(fasta="x.unique.fasta")

# align these sequences
align.seqs(fasta="x.unique.fasta", template="silva.bacteria.fasta", processors=4)
```

Get the data from the course webpage