2013 - BMMB 597D: Analyzing Next Generation Sequencing Data

Week 13, Lecture 25

István Albert

Biochemistry and Molecular Biology and Bioinformatics Consulting Center

Penn State
Introduction to metagenomics

• Genetic material recovered from environmental samples

• Note that no living organism exists on its own

• Human body is estimated to contain 10 trillion cells

• The human body contains more bacteria than cells (estimates of 10:1) ➔ 100 trillion bacteria! (is that really true? in class discussion)
Metagenomics approaches

• metagenome assembly → reconstitute entire genomes from randomly fragmented short reads:
  – many (mostly) unknown genomes
  – genomes with various abundances
  – sequences with systematic errors
  – more data more unreliable results?

How may genomes could there be in the first place?

Some say: over 1 billion!

Caveat emptor: 1 billion sounds like a really big number

and we just saw an example of where big numbers come from
Simpson’s paradox

A famous example: berkley gender bias lawsuit

<table>
<thead>
<tr>
<th></th>
<th>Applicants</th>
<th>Admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>8442</td>
<td>44%</td>
</tr>
<tr>
<td>Women</td>
<td>4321</td>
<td>35%</td>
</tr>
</tbody>
</table>

Total admission rates seem to statistically favor **men**.

<table>
<thead>
<tr>
<th>Department</th>
<th>Men</th>
<th></th>
<th>Women</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>Admitted</td>
<td>Applicants</td>
<td>Admitted</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>62%</td>
<td>108</td>
<td>82%</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63%</td>
<td>25</td>
<td>68%</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37%</td>
<td>593</td>
<td>34%</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33%</td>
<td>375</td>
<td>35%</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28%</td>
<td>393</td>
<td>24%</td>
</tr>
<tr>
<td>F</td>
<td>272</td>
<td>6%</td>
<td>341</td>
<td>7%</td>
</tr>
</tbody>
</table>

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.
Phylotyping

- Map the reads against all known genetic sequences

- Find exact or partial (homologous) matches, characterize sample by the known genomes

- There are not that many fully sequenced bacteria - thousands? But there is data on lots of partially sequenced ones – millions.

- The most commonly used tool is BLAST
OTU based approaches

• OTU – operational taxonomical units, most common approach, no taxonomy required:

  • sequences are characterized relative to their similarity to one another

  • groups are formed based on similarity thresholds
16S RNA

- the gene 16S rRNA gene is highly conserved between different species of bacteria and archaea and:
  - parts of these genes are identical in just about all bacteria → universal primers
  - other sections of the 16S rRNA gene contain species-specific signature
  - LSU, SSU → large subunit and small subunit

rNRA → decodes mRNA into aminoacids → essential to life
16S rRNA sequencing

- isolate only the 16S rRNA genes, sequence only these

Pros: require far less coverage, allows, characterize population

Cons: we don’t know what the unknown bacteria are like other then the region that we sequenced
Strenghts and weaknesses

• Phylotyping → shifts/biases the view towards known sequences, has to rely on external taxonomy

• OTU → is far more abstract, it cannot tell what is actually there, grouping is subjective

Most often people use a combination of all techniques
Complication: competing standards

- Multiple sources of bacterial taxonomical classification
Competing 16s rRNA taxonomical databases

- **silva** – comprehensive ribosomal RNA database → ARB, Germany

- **rdp** → ribosomal database project → Michigan State, USA

- **greengenes** → comprehensive 16S rRNA gene sequence alignment

- also can download one from GenBank

Each of these resources distributes a slightly/or not so slightly different taxonomy → Silva has 9 levels
Some tools mandate the use of the NCBI taxonomy

How to reference the NCBI taxonomy database

The NCBI taxonomy database is not a primary source for taxonomic or phylogenetic information. Furthermore, the database does not follow a single taxonomic treatise but rather attempts to incorporate phylogenetic and taxonomic knowledge from a variety of sources, including the published literature, web-based databases, and the advice of sequence submitters and outside taxonomy experts. Consequently, the NCBI taxonomy database is not a phylogenetic or taxonomic authority and should not be cited as such.
Commonly used 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**

• **MEGAN** Metagenome Analyzer → visualization package
Lot of authors, published in Nature Methods, 2010
Has a Google Group forum that seem very active
it is a Python based glue that connects many independent tools
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 advance methods and techniques!

• You can perform a full analysis using only mothur

• Excellent documentation (though some aspects refer to previous mothur releases so may be out of sync)
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
Download and install the RPD classifier

$ java -Xmx1g -jar ~/src/rdpClassifier_2.6/dist/classifier.jar
USAGE: ClassifierMain <subcommand> <subcommand args >
default command is classify
classify - classify one or multiple samples
libcompare - compare two samples
merge - merge multiple classification result files to create a new hier_out file
train - retrain classifier
loot - leave-one-out accuracy testing
crossvalidate - cross validate accuracy testing

yalbert@porthos ~/work/lec25
$ java -Xmx1g -jar ~/src/rdpClassifier_2.6/dist/classifier.jar classify meta.fa -f fixrank -o sequences.txt -h meta.fa.rdp.txt


**Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.**

Wang Q, Garrity GM, Tiedje JM, Cole JR.
Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824, USA.
### Classification file

<table>
<thead>
<tr>
<th>taxid</th>
<th>lineage</th>
<th>name</th>
<th>rank</th>
<th>meta.fa</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 null</td>
<td>Root; rootrank; Bacteria; dom Bacteria</td>
<td>null</td>
<td>no rank</td>
<td>4767</td>
</tr>
<tr>
<td>1</td>
<td>Root; rootrank; Bacteria; dom Acidobacteria</td>
<td>domain</td>
<td></td>
<td>4767</td>
</tr>
<tr>
<td>2195</td>
<td>Root; rootrank; Bacteria; dom Acidobacteria_Gp2</td>
<td>phylum</td>
<td></td>
<td>1732</td>
</tr>
<tr>
<td>2206</td>
<td>Root; rootrank; Bacteria; dom Acidobacteria_Gp2</td>
<td>class</td>
<td></td>
<td>387</td>
</tr>
<tr>
<td>2207</td>
<td>Root; rootrank; Bacteria; dom Gp2</td>
<td>genus</td>
<td></td>
<td>387</td>
</tr>
<tr>
<td>2208</td>
<td>Root; rootrank; Bacteria; dom Acidobacteria_Gp3</td>
<td>class</td>
<td></td>
<td>135</td>
</tr>
<tr>
<td>2209</td>
<td>Root; rootrank; Bacteria; dom Gp3</td>
<td>genus</td>
<td></td>
<td>133</td>
</tr>
<tr>
<td>2204</td>
<td>Root; rootrank; Bacteria; dom Acidobacteria_Gp1</td>
<td>class</td>
<td></td>
<td>864</td>
</tr>
<tr>
<td>2205</td>
<td>Root; rootrank; Bacteria; dom Gp1</td>
<td>genus</td>
<td></td>
<td>719</td>
</tr>
<tr>
<td>2256</td>
<td>Root; rootrank; Bacteria; dom Granulicella</td>
<td>genus</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>2255</td>
<td>Root; rootrank; Bacteria; dom Edaphobacter</td>
<td>genus</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>2216</td>
<td>Root; rootrank; Bacteria; dom Acidobacteria_Gp7</td>
<td>class</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>2217</td>
<td>Root; rootrank; Bacteria; dom Gp7</td>
<td>genus</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>-2205</td>
<td>Root; rootrank; Bacteria; dom unclassified_Acidobacteria_Gp1</td>
<td>genus</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>2214</td>
<td>Root; rootrank; Bacteria; dom Acidobacteria_Gp6</td>
<td>class</td>
<td></td>
<td>56</td>
</tr>
<tr>
<td>2215</td>
<td>Root; rootrank; Bacteria; dom Gp6</td>
<td>genus</td>
<td></td>
<td>56</td>
</tr>
<tr>
<td>2210</td>
<td>Root; rootrank; Bacteria; dom Acidobacteria_Gp4</td>
<td>class</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>2211</td>
<td>Root; rootrank; Bacteria; dom Gp4</td>
<td>genus</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>2212</td>
<td>Root; rootrank; Bacteria; dom Acidobacteria_Gp5</td>
<td>class</td>
<td></td>
<td>67</td>
</tr>
<tr>
<td>2213</td>
<td>Root; rootrank; Bacteria; dom Gp5</td>
<td>genus</td>
<td></td>
<td>67</td>
</tr>
</tbody>
</table>
MEGAN: Metagenome Analyzer
MEGAN is the best visualizer
Homework 25

- Using the RPD multiclassifier and the meta.fa data from the course website answer the following:

  - Which phylum is the most abundant?
  - Which genus is the most abundant?
  - What percent of reads are unclassified even at domain level?