Phylogeny
- based on whole genome data

Johanne Ahrenfeldt
PhD student
Overview

• What is Phylogeny and what can it be used for
• Single Nucleotide Polymorphism (SNP) methods
  - snpTree and CSI Phylogeny
• Nucleotide Differences
  - NDtree
• Controlled Evolution study
• What services for which data
What are phylogenetic trees

- Trees are traditionally made using aligned sequences of single genes or proteins
- Whole genome data may be used to create trees based on
  - SNP calling
  - K-mer overlap
What is a SNP

- A Single Nucleotide Polymorphism (SNP) is a DNA sequence variation occurring commonly* within a population (e.g. 1%) in which a Single Nucleotide — A, T, C or G — in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes.
What is phylogeny used for

• Classify taxonomy – The classic use
• Outbreak detection – Increasing with WGS data
How does it work

Strain A  AT\textbf{TCA G T A}
Strain B  AT\textbf{GCA G T C}
Strain C  AT\textbf{GCA AT C}
Strain D  AT\textbf{TCA G T C}
Construct distance matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Make Tree

Strain A   ATTCAGTA
Strain B   ATGCGAGTC
Strain C   ATGCAATC
Strain D   ATTCAGTC

A   B   C   D
A  0   2   3   1
B  2   0   1   1
C  3   1   0   2
D  1   1   2   0
snpTree

- First online webserver for constructing phylogenetic trees based on whole genome sequencing

snpTree flow

A

1. Raw reads
2. Pre-processing
3. Reads mapping (using BWA)
4. Identify SNPs (using SAMtools)
5. SNPs filtering (using SAMtools)
6. SNPs tree construction (using Fasttree)

B

1. Assembled genomes
2. Reference genome alignment (using Nucmer)
3. Identify SNPs (using show-snps from MUMmer)
4. SNPs filtering (using show-snps from MUMmer)
5. SNPs tree construction (using Fasttree)
CSI Phylogeny

• SNP identification same as snpTree

• Strict sorting of SNPs
  – Depth
  – Relative depth
  – Distance between SNPs
  – SNP quality
  – Read mapping quality

CSI Phylogeny

- Requires all SNPs to be significant
  - Z-score higher than 1.96 for all SNPs

\[ Z = \frac{X - Y}{\sqrt{X+Y}} \]

- X is the number of reads, with the most common nucleotide at that position, and Y the number of reads with any other nucleotide.
**Nucleotide calling**

- A different approach where the main distinction is not between if a SNP should be called or not, but between whether or not there is solid evidence for what nucleotide should be called.
Ndtree

Simple mapping approach

• Cuts all reads into Kmers
• Maps all Kmers to reference genome
• Makes ungapped consensus sequences of equal lengths
Nucleotide calling

• When all reads have been mapped the significance of the base call at each position was evaluated by calculating the number of reads $X$ having the most common nucleotide at that position, and the number of reads $Y$ supporting other nucleotides.

• A Z-score threshold was calculated as:

$$Z = \frac{X - Y}{\sqrt{X+Y}} > 1.96 \text{ (or 3.29)}$$

• $>90\%$ of reads supporting the same base
Count nucleotide differences

- **Method 1:** Each pair of sequences was compared and the number of nucleotide differences in positions called in **all** sequences was counted.
  - More accurate (Z=1.96 is used as threshold)

- **Method 2:** Each pair of sequences was compared and the number of nucleotide differences in positions called in **both** sequences was counted.
  - More robust (Z=3.29 is used as threshold)
NDtree

Uses two different algorithms to make two different trees

• UPGMA
• Neighbor Joining

Both algorithms are part of the Neighbor program package and makes trees from distance matrices
Controlled Evolution study

Choose colony
Grow for 8 h
Plate out Grow for 8 h
Choose colonies
Grow for 16 h
Choose colonies
Plate out Grow for 8 h
Grow for 16 h

Day 1
Day 2
Day 3
... Day

128x

Johanne Ahrenfeldt, Master thesis.
<table>
<thead>
<tr>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>....</th>
<th>Day 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1_1</td>
<td>1_1_1</td>
<td>1_1_1</td>
<td>a_1_1_1_1_1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1_1</td>
<td>1_1_1</td>
<td>1_1_2</td>
<td>b_1_1_1_1_2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1_2</td>
<td>1_1_2</td>
<td>1_1_2</td>
<td>c_1_1_1_2_1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1_2</td>
<td>1_1_2</td>
<td>1_1_2</td>
<td>d_1_1_1_2_2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1_2</td>
<td>1_2_1</td>
<td>1_2_1</td>
<td>e_1_1_2_1_1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1_2</td>
<td>1_2_1</td>
<td>1_2_1</td>
<td>f_1_1_2_1_2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1_2</td>
<td>1_2_1</td>
<td>1_2_1</td>
<td>g_1_1_2_2_1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1_2</td>
<td>1_2_1</td>
<td>1_2_1</td>
<td>h_1_1_2_2_2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1_2</td>
<td>1_2_1</td>
<td>1_2_1</td>
<td>i_1_2_1_1_1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1_2</td>
<td>1_2_1</td>
<td>1_2_1</td>
<td>j_1_2_1_1_2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1_2</td>
<td>1_2_2</td>
<td>1_2_2</td>
<td>k_1_2_1_2_1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1_2</td>
<td>1_2_2</td>
<td>1_2_2</td>
<td>l_1_2_1_2_2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1_2</td>
<td>1_2_2</td>
<td>1_2_2</td>
<td>m_1_2_2_1_1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1_2</td>
<td>1_2_2</td>
<td>1_2_2</td>
<td>n_1_2_2_1_2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1_2</td>
<td>1_2_2</td>
<td>1_2_2</td>
<td>o_1_2_2_2_1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Phylogenetic tree using NDtree (UPGMA)
Phylogenetic tree using NDtree (Neighbor Joining)
UPGMA vs. Neighbor Joining

- UPGMA works well when samples have been taken the same time

- Neighbor joining is better when samples have been taken at different times
So... What should I use when?

• CSI Phylogeny is advantageous to use when you expect the differences between the samples to be larger than 5-10 mutations.

• NDtree on the other hand is able to find these small differences, but may not be strict enough to handle very large differences.
Thank you for listening

• Questions?