Whole Genome based Phylogeny

Johanne Ahrenfeldt
PhD student
DTU Bioinformatics
Short about me

Johanne Ahrenfeldt
johah@cbs.dtu.dk

• PhD student at DTU Bioinformatics
  – Whole Genome based Phylogeny
• Graduate Engineer in Systems Biology and Bioinformatics from Technical University of Denmark
• Working in the CGE project since 2012 – started as a student helper
Overview

• What is Phylogeny
• SNP methods
  – CSI Phylogeny
• Nucleotide Differences
  – NDtree
• Controlled Evolution study
• Good advice
What is phylogeny?

• Early phylogeny
  – Classification
  – Based on phenotypes

• Current phylogeny
  – Based on genotypes
  – DNA mutations as basis for evolution
Classification

Carl Linnaeus 1707-1778

Hierarchical system

Kingdom
Phylum
Class
Order
Family
Genus
Species
Classification depicted as a tree.

\[ \text{Order} \rightarrow \text{Family} \rightarrow \text{Genus} \rightarrow \text{Species} \]
Classification depicted as a tree

Species  Genus  Family  Order  Class
DNA mutations as basis for evolution

DTU Bioinformatics
Department of Bio and Health Informatics
What are phylogenetic trees

- Phylogenetic trees are a visual representation of the genetic relationship between species
- Think of them as family trees
- Phylogeny can also be represented by distance matrices
What are phylogenetic trees

- Trees were traditionally made using aligned sequences of single genes or proteins
- Whole genome data can be used to create trees based on
  - SNP calling
  - K-mer overlap
  - Alignment of genomes
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.
How does it work

Strain A       ATTCAGTAGT
Strain B       ATGCAAGTTGA
Strain C       ATGCAATTTGT
Strain D       ATCCATTAGC
Construct distance matrix

<table>
<thead>
<tr>
<th>Strain A</th>
<th>ATCTAGTGGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain B</td>
<td>ATGCGTTGA</td>
</tr>
<tr>
<td>Strain C</td>
<td>ATGCATTGTA</td>
</tr>
<tr>
<td>Strain D</td>
<td>ATCCATTAGC</td>
</tr>
</tbody>
</table>

<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></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Make Tree

Strain A  AT\text{CAGTAGT}  
Strain B  AT\text{GCAGTTGA}  
Strain C  AT\text{GCAATTGT}  
Strain D  AT\text{CCATTAGC}  

<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>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

DTU Bioinformatics
Department of Bio and Health Informatics
How to read phylogenetic trees

$$A = (B + D) = (B + C + E + G) = (B + C + F + H)$$

$$D = (C + E + G) = (C + F + H)$$

[Diagram of phylogenetic tree with labels A, B, C, D, E, F, G, H, and a human figure at the root.]
How to read phylogenetic trees

a

b

DOI 10.1007/s12052-008-0035-x
What is phylogeny used for

- Classify taxonomy – The classic use
- Outbreak detection – Increasing with WGS data
What is phylogeny used for

• Cholera outbreak in Haiti 2010
• Listeria outbreak 2014

Case story

- *Vibrio Cholerae* outbreak in Haiti followed the 2010 earthquake
- Rumors said that the outbreak may have come from Nepal, travelling along with UN soldiers from Nepal
- No proof had been given of this until the Hendriksen *et al.* paper in 2011

Case story

• Data
  – 24 recent *V. cholerae* strains from Nepal
  – 10 previously sequenced *V. cholerae* isolates, including 3 from the Haitian outbreak

• Analysis
  – Antimicrobial susceptibility testing
  – PFGE (pulsed-field gel electrophoresis) to analyze for genetic relatedness
  – Whole genome sequencing, SNP identification and phylogenetic analysis
## Case story - Results

<table>
<thead>
<tr>
<th>Resistance profile</th>
<th>Susceptible</th>
<th>Decreased susceptibility</th>
<th>Resistant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nepalese strains</td>
<td>Tetracycline</td>
<td>Ciprofloxacin</td>
<td>Trimethoprim, Sulfamethoxazole, Nalidixic</td>
</tr>
<tr>
<td>Hendriksen et al. 2011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haitian outbreak strains</td>
<td>Tetracycline</td>
<td>Ciprofloxacin</td>
<td>Trimethoprim, Sulfamethoxazole, Nalidixic</td>
</tr>
<tr>
<td>Centers for Disease Control and Prevention, 2010</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Case story - Results

- Pulsed-field gel electrophoresis (PFG)E
  - Nepalese isolates divided in 4 groups
  - Most common Haitian type in same group as four Nepalese strains
Case story - Results

FIG 1

Genetic relationships among *V. cholerae* isolates from Nepal and Haiti. A single maximum parsimony tree was reconstructed using 752 SNPs from 34 whole-genome sequences. There were 184 parsimony-informative SNPs, of which 6 were homoplastic, resulting in a CI of 0.97 (excluding uninformative characters). The branch lengths are labeled in red, and for branches affected by homoplasy, minimum and maximum branch lengths are designated. Members of SNP genotypic group V (16) are indicated. SNP differences among the three most closely related Nepali groups and the Haitian group are shown and characterized in Table S1 in the supplemental material.

TABLE 1

Different point mutations observed among the three sequenced isolates from the Haiti outbreak and the three most closely related isolates from Nepal

<table>
<thead>
<tr>
<th>Chromosome Position</th>
<th>Nucleotide or amino acid in:</th>
<th>Reference strain</th>
<th>Haitian isolate</th>
<th>Nepalese isolate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1786</td>
<td>1792</td>
<td>1798</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>2387016</td>
<td>C</td>
<td>CCCT T</td>
<td>T T</td>
<td>Gly Gly Gly Arg</td>
</tr>
<tr>
<td>1090536</td>
<td>T</td>
<td>TTTT T</td>
<td>G</td>
<td>Ile</td>
</tr>
<tr>
<td>962762</td>
<td>C</td>
<td>CCCT C</td>
<td>C C</td>
<td>Ala Ala Ala Ala</td>
</tr>
</tbody>
</table>

° The reference strain is *Vibrio cholerae* O1 biovar El Tor strain N16961 (Bangladesh 1971). The NCBI reference sequences or accession numbers are NC_002505 for chromosome I and NC_002506 for chromosome II.
10 minutes break!
snpTree

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

snpTree flow

A

Raw reads

Pre-processing

Reads mapping (using BWA)

Identify SNPs (using SAMtools)

SNPs filtering (using SAMtools)

SNPs tree construction (using Fasttree)

B

Assembled genomes

Reference genome alignment (using Nucmer)

Identify SNPs (using show-snps from MUMmer)

SNPs filtering (using show-snps from MUMmer)

SNPs tree construction (using Fasttree)
CSI Phylogeny

https://cge.cbs.dtu.dk/services/CSIPhylogeny/

• 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.
CSI Phylogeny

Output

Tree build by FastTree algorithm, in Newick format
  • Branch lengths is substitutions per site at the variable sites

Matrix of SNP pair counts in text (.txt) format
  • Diagonal SNP matrix
CSI Phylogeny

Download the filtered SNP calls in Variant Calling Format (VCF):
Note: VCF files are compressed with gzip.

Download matrix of SNP pair counts:
Download matrix as: TXT  EPS

Download SNP alignment: FASTA

Percentage of reference genome covered by all isolates: 95.6684818250054
4440598 positions was found in all analyzed genomes.
Size of reference genome: 4641652

Below is listed the number of positions that are shared and trusted between each isolate and the reference genome.

<table>
<thead>
<tr>
<th>File</th>
<th>Valid positions</th>
<th>Pct. of reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_1_2_2_1_1_2_1_R1.ignored_snps 4448690</td>
<td>95.8428163076422</td>
<td></td>
</tr>
<tr>
<td>1_2_1_1_2_1_2_2_R1.ignored_snps 4450004</td>
<td>95.8711251942196</td>
<td></td>
</tr>
</tbody>
</table>

DTU Bioinformatics
Department of Bio and Health Informatics
NDtree

https://cge.cbs.dtu.dk/services/NDtree/

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 the nucleotide at the given position.

NDtree

Simple mapping approach

• Cuts all reads into K-mers
• Maps all K-mers to reference genome
• Makes an ungapped consensus sequences of equal lengths
Mapping

K-mers

Reference genome

Consensus sequence

Reference genome
Genome 1
Genome 2
Genome 3
Genome 4
Genome 5
Genome 6
NDtree

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 is calculated

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

$>90\%$ of reads supporting the same base
NDtree

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)
Method 1 – all called

- Significant positions in Genome 1
- Significant positions in Genome 2
- Significant positions in Genome 3
- Positions used for phylogeny

Method 2 – pairwise significance

- Significant positions in Genome 1
- Significant positions in Genome 2
- Significant positions in Genome 3
- Positions used between 1 and 2
- Positions used between 1 and 3
- Positions used between 2 and 3
NDtree

Uses two different algorithms to make two different trees
• UPGMA
• Neighbor Joining

Both algorithms are part of the PHYLIP Neighbor program package and make trees from distance matrices
UPGMA vs. Neighbor Joining

• UPGMA works when samples have been taken the same time

• Neighbor Joining is better when samples have been taken at different times
NDtree

Output

- **distance.txt**: Distance matrix - tab separated
- **dist.mat**: Distance matrix - PHYLIP format
- **tree.nj.newick**: Neighbor Joining tree - Newick format
  - Branch lengths is number of Nucleotide Differences
- **tree.upgma.newick**: UPGMA tree – Newick format
  - Branch lengths is number of Nucleotide Differences
For each 8 hour culture a sample was saved for DNA sequencing.

## Naming the descendants

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td></td>
<td></td>
<td>S1</td>
<td></td>
<td>S1111</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S11</td>
<td></td>
<td>S1112</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S11</td>
<td>S1121</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S112</td>
<td>S1122</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S12</td>
<td></td>
<td>S1211</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S121</td>
<td></td>
<td>S1212</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S122</td>
<td>S1221</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S1222</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2</td>
<td>S21</td>
<td></td>
<td>S2111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S21</td>
<td>S211</td>
<td></td>
<td>S2112</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S212</td>
<td>S2121</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S2122</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td></td>
<td>S22</td>
<td></td>
<td>S2211</td>
</tr>
<tr>
<td></td>
<td>S22</td>
<td></td>
<td>S221</td>
<td></td>
<td>S2212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S222</td>
<td>S222</td>
<td></td>
<td>S2221</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S222</td>
<td>S2222</td>
</tr>
</tbody>
</table>
Mutations
Phylogenetic tree using NDtree (UPGMA)
Phylogenetic tree using NDtree (Neighbor Joining)
UPGMA vs. Neighbor Joining

- UPGMA works when samples have been taken the same time

- Neighbor Joining is better when samples have been taken at different times
CSI Phylogeny – Pruning disabled
So... What should I use when?

CSI Phylogeny
• Has very good statistics and a good graphical overview.
• Advantageous to use when you expect the differences between the isolates to be larger than 5-10 mutations.
• Is faster

NDtree
• Is able to find very small differences.
• Does not take recombination into consideration.
• Works best on raw reads. If given assembled genomes, it simulates reads.
Choosing a reference genome

For comparison of very closely related isolates, a better level of detail is given by using a closely related reference genome.
What defines an outbreak

• We can’t tell for certain
• It depends on the species
• But a rule of thump is:
  – Within 10 SNPs it is definitely an outbreak
  – Within 30 SNPs it might be an outbreak
  – Above 60 SNPs it is most likely not an outbreak
Thank you for listening

• Questions?