Project II: Bacterial Phylogeny

Learning Objectives

The student will be able to

- convert sequence data from Excel to FASTA file format (optional)
- align multiple DNA sequences using a general purpose multiple sequence alignment program for DNA sequences (ClustalW2)
- interpret a phylogenetic tree

Knowledge and Skills

- FASTA format
- Online resources for multiple sequence alignment and tree building
- Phylogenetic trees

Prerequisites

- Optional: Familiarity with Matlab

Source: Lieberman, T.D. *et al.* Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genetics* 43: 1174-1176 doi:10.1038/ng.1011 (http://www.nature.com/ng/journal/v43/n12/full/ng.997.html)

A retrospective study of 112 isolates from 14 individuals with cystic fibrosis collected during an outbreak of *Burkholderia dolosa* over 16 years identifies genetic mutations.

Task: Infer the likely transmission network between individuals and between lung and blood isolates within individuals based on the bacterial phylogeny of the 112 isolates.

Citation: Neuhauser, C. Project II: Determining Mutation Bacterial Phylogeny.
Created: January 1, 2012 **Revisions:**

FASTA Format

http://en.wikipedia.org/wiki/FASTA_format

The FASTA format is a text-based format for DNA or protein sequences that can be read by many software packages. Each sequence begins with a single-line starting with the “>” symbol to mark the start of a new sequence, followed by the description of the sequence, and, in a new line, the sequence data. There is no space between the “>” symbol and the sequence description. The end of the sequence is indicated by the start symbol “>” of the next sequence. The sequence of the first patient in the sample in FASTA format begins as follows

>A-0-0
CGCAATGTCGTCTTCTCCGGCTCTTTGGGCGGCGCCTACTTCGACAGCGGGTGGCCTCCTCTGACGACGTCCCGAG
CGCAGAAG

Optional: Converting sequence data stored in Excel to a FASTA format text file

Matlab provides a command to convert sequence data from Excel to FASTA format. The first step is to read the data from a spreadsheet. We store the names of the sequences and the sequence data in two separate sheets. The first sheet contains the sequence names in the first row; the second sheet contains the sequence data, one sequence per column, The following Matlab script stores the sequence data and their names in the file ‘pba.’

```
clear
[num,txt,raw] = xlsread('pbeData.xlsx','DNASequence');
[num1,txt1,raw1] = xlsread('pbeData.xlsx','Description');

save pba
```

The variable txt1 has the names of all sequences in the first row. The variable txt has the sequence data, one sequence per column. The following scripts converts the sequence data into the FASTA format.

```
clear
load pba

%We load 'pba', which has all the sequence information as text in matrix
%form. The variable txt is a 511x115 matrix where each cell of the matrix
%has a single letter of one of the sequences. Each column is a sequence
%The variable txt1 has the labels in the first row.

lseq=size(txt,1); %Length of the sequences
nseq=size(txt,2); %Number of sequences

%We need to convert that sequence data in fasta file format. For this, we
```

```
%string the letters together and add the headers.

txtst=txt(1,:);
for i=2:lseq
    txtst=strcat(txtst,txt(i,:));
end
pbast=txtst';
pbatxt=txt1';

for i=1:nseq
    seda(i).Sequence=pbast{i,1};
    seda(i).Header=pbatxt{i,1};
end

%The sequences are stored in fasta format in the 'burkisolates.txt' file
fastawrite('burkisolates.txt', seda)
```

The FASTA formatted sequence data is stored in the text file burkisolates.txt. The first three sequences are the reference sequence AU0158, the outgroup X-3-7, and the Last common ancestor. The remaining sequences are the 112 isolates.

Bacterial Phylogeny

The sequence data is in the burkisolates.txt file in FASTA format. To build a tree, we need the sequence data from the 112 isolates and the outgroup.

Copy the burkisolates.txt into a new text file. Call the file burkphylo.txt. Delete the sequence data of the reference sequence and the last common ancestor. We will use ClustalW2 to build the tree. ClustalW2 aligns multiple sequences. Go to http://www.ebi.ac.uk/Tools/msa/clustalw2/ and follow the steps. Make sure you select DNA in Step 1. Use the default options in Step 2 and Step 3. Submit your job.

When your job is ready, you will see the alignment in the same browser window. The Guide Tree tab on top shows the tree (you need to scroll down until the Phylogram).

Help: http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny/help/faq.html#15

Task 1:

Investigate the sequence of transmission among the individuals. Which of the isolates cluster according to subject?

Task 2:

Blood is usually a sterile medium, that is, bacteria are not found in the blood. Bacteremia is a condition that indicates bacteria in the blood. Since subjects K, N, and H have isolates from both the airways and the bloodstream, they suffered from bacteremia. In which individuals can you find evidence of multiple transmissions from the lungs to the bloodstream?

Citation: Neuhauser, C. Project II: Determining Mutation Bacterial Phylogeny.
Created: January 1, 2012 **Revisions:**