PLPTH813, Homework 4

Due on 4/4/2019

The purpose of this homework is to practice variant discovery and tree construction.

**Problem 1.**

The infection of Escherichia coli O157:H7 may lead to hemorrhagic diarrhea and kidney failure. The E. coli outbreak occurred in 1993 was due to the contamination of E.coli O157:H7 bacterium from undercooked beef patties in hamburgers. In this problem, we need to examine the number SNP/INDEL variants of an O157:H7 strain relative to an E.coli reference genome M1655.

Data:

Sequencing data can be copied from the following directory at Beocat:

**/homes/liu3zhen/teaching/datasets/HW04**

These data were randomly sampling paired-end reads from the Genbank accession of SRR957857.

The reference genome is the same as the one we used in our lab.

Run the SNP/INDEL discovery analysis (alignment, data formatting such as SAM to BAM, and GATK variant discovery) using GATK and report the procedure, the scripts and the number of SNPs/INDELs you identified.

**Problem 2.**

Below are four DNA sequences that will be used to construct a phylogenetic tree.

A CATGCATGAG

B CAAGCACGAG

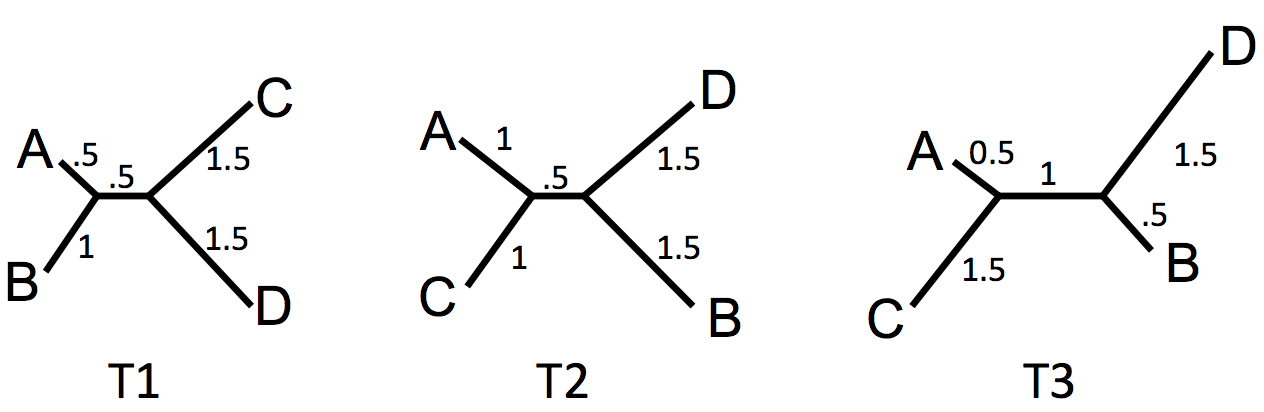
C CATGCTTGCG

D CAAGAATGCG

1) Determine the number of difference between each pair of two sequences and fill the number in the following matrix.

|  |  |  |  |
| --- | --- | --- | --- |
|  | B | C | D |
| A |  |  |  |
| B |  |  |  |
| C |  |  |  |

2) Refer to the lecture for the algorithms of “least squares” and “minimum evolution” and select the “best” tree from the trees that were illustrated below. Describe the procedure and justify the reason for your selection.



**Problem 3.**

seq1: ATC

seq2: CTG

seq3: ATG

seq4: ACC

A phylogenetic tree needs to be constructed for these four DNA sequences. The algorithm of **maximum parsimony** will be used for the tree construction. Write down the nucleotide difference (e.g., 0, 1, 2 …) for each edge in the following tree and determine the tree score. Describe which tree you will choose and justify your choice.

ATC ACC

\ /

\ /

ATG --- CTG Tree I score:

/ \

/ \

ATG CTG

ATC ATG

\ /

\ /

ATC --- ATG Tree II score:

/ \

/ \

ACC CTG

ATC ATG

\ /

\ /

ATG --- ATG Tree III score:

/ \

/ \

CTG ACC