Build Phylogenetic Tree

This project requires you to implement the neighbor joining algorithm and the majority-rule tree algorithm.

You are asked to create 3 programs NJ, Bootstrap, and Mconsensus.

- NJ generates the neighbor joining tree given a set of aligned sequences in “seq.aln”.
- Given a set of aligned sequences “seq.aln”, Bootstrap randomly generates *k* set of aligned sequences. For the *k* set of aligned sequences, it generates *k* neighbor joining trees. (The seed of the random number generator is assumed to be *seed*.)
- Given *k* trees in “trees.dnd”, Mconsensus computes the majority-rule consensus tree.

If you are using java, the command line is as follows.

```
java NJ seq.aln > tree.dnd
java Bootstrap seq.aln seed k > trees.dnd
java Mconsensus trees.dnd > tree.dnd
```

If you are using C, the command line is as follows.

```
NJ seq.aln  > tree.dnd
Bootstrap seq.aln seed trials > trees.dnd
Mconsensus trees.dnd > tree.dnd
```

Distance between two species

For any two aligned sequences, let ℓ be the length after removing the gaps and h be the aligned characters which are the same. Then, the distance of the two aligned sequences is $-\ln(1-h/\ell)$.

Bootstrapping

Given a set of aligned sequences of length m, bootstrapping randomly selects m aligned columns to form a new set of aligned sequences.

FILE FORMAT

Please refer to the two test datasets for the format of the aligned sequences seq.aln. (This is the format used by clustalW. Please also read ClustalW_aln_format.txt.)

The tree.dnd is written in the Phylip tree format. (Please read Phylip_dnd_tree_format.txt for the detail of the format.)

Note that you can visualize your tree using the software Treeview (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

Detail of the assignment

You are required to write your program using either C or java. (For C, please make sure you are writing ANSI C. Otherwise, I may be unable to compile your program.)

The program names and the file format must strictly follow what we stated above.

Testing data

You are given two sets of testing data.

The first testing dataset is test1.aln. After executing the following command lines, we may get the corresponding sample output files test1_trees.dnd and test1_consensus.dnd. (Note: 1234 is the seed for the random number generater. Different seeds will give different output.)

```
java Bootstrap test1.aln 1234 3 > test1_trees.dnd
java Mconsensus trees.dnd > test1_consensus.dnd
```

The second testing dataset is Seq_HIV4.aln. After executing the following command lines, we may get the corresponding sample output file Seq_HIV4.dnd.

```
java Bootstrap Seq_HIV4.aln 1234 100 > trees.dnd
java Mconsensus trees.dnd > Seq_HIV4.dnd
```

Submission

Please submit your project and handin a document which describes the detail of your implementation and the trees you generated using the two testing datasets (assume the number of bootstrapping trials is 100).

Please make sure your programs can handle big dataset and is efficient enough.