

Algorithms in Bioinformatics I, WS2002/3

Assignment sheet # 9

Daniel Huson

December 9, 2002

The goal of the following assignments is to write a simulation package from testing tree reconstruction programs.

To solve the problems below, please download the following file:

www-ab.informatik.uni-tuebingen.de/teaching/ws02/abi1/programs/program11.zip.

The download contains a number of java files:

- The file called `JCSimulator.java` contains the frame-work for a program that takes as input a rooted tree with edge lengths, a sequence length L and a mutation rate u . In the first assignment you are requested to implement the simulator and the goal is to produce sequences and Hamming distances as output.
- The file `Distances.java` implements a distance matrix. (However, this version does not contain the neighbor-joining algorithm.)
- The file `Characters.java` implements a list of aligned sequences. In the second assignment you are requested to add a method that computes the Hamming distances for the sequences and to add a method that computes the Jukes-Cantor transformed distances for the sequences.
- The file `CharactersToDistances.java` contains a program that reads as input aligned sequences and produces as output their Hamming distances or Jukes-Cantor transformed distances. Please add your name to this program, without modifying anything else.
- The file `PhylogeneticTree.java` that provides all necessary functionality of a phylogenetic tree. No modification of this file is necessary.
- Additionally, the download also contains all necessary `jloda` classes and their documentation.

Please run the program on all supplied data sets.

1 Simulating sequences using the Jukes-Cantor method (6 points)

Write a program `JCSimulator.java` that takes as input a rooted tree (with branch lengths), a sequence length L and a mutation rate u . Output is a system of aligned sequences of length L generated under the Jukes-Cantor model of evolution.

2 Computing Hamming and Jukes-Cantor distances (2 points)

Write a program `CharactersToDistances.java` that reads a list of aligned sequences and prints out their Hamming distances and Jukes-Cantor transformed distances. To do this, implement two new methods `getHamming` and `getJukesCantor` in `Characters.java`.

3 Run neighbor-joining on simulated data (2 points)

For each of the input files, choose three different reasonable values for L and u , generate sequences and run the neighbor-joining program on the Hamming and Jukes-Cantor distances. Compare the input trees with the output trees. How does the performance depend on the sequence length and mutation rate?

Due by 10am, Monday, 16 Dec 2002