

Algorithms in Bioinformatics I, WS2006/07

Assignment sheet # 13

Christian Rausch

January 22, 2007

1 Average Hamming distance between random sequences (1 point)

What is the average Hamming distance between two random, gap-free DNA sequences of the same length (with all nucleotides equally probable)?

2 Runtime complexity of UPGMA (1 point)

State and prove the time complexity of a naïve implementation of UPGMA as a function of the number of taxa.

3 Speed-up of UPGMA (3 points)

Modify the UPGMA algorithm so that its time complexity becomes quadratic function of the number of taxa. (Hint: maintain additional information to find the minimal entry in the distance matrix faster.) UPGMA is sometimes *called* a linear time algorithm, why?

4 Implementation of Neighbor-Joining (5 points)

Write a program that reads in a distance matrix and then computes and displays the Neighbor-Joining tree. To do so, please download:

<http://www-ab.informatik.uni-tuebingen.de/teaching/ws06/albi1/assign/java/java13.zip>,

and modify the file `NeighborJoining.java`, located in the package `albi.phylo`.

Note: Exercise 3.4 of assignment sheet 12 and all exercises of this sheet are due by 10am, Monday, 29 Jan 2007.

Assignment sheet 13 will thus have 13 total points.