

Algorithms in Bioinformatics I, WS2006/7

Assignment sheet # 14

Tobias Dezulian

January 29, 2007

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

<http://www-ab.informatik.uni-tuebingen.de/teaching/ws06/albi1/assign/java/java14.zip>

1 The Jukes-Cantor distance transformation (2 points)

Please derive the *Jukes-Cantor distance transformation* (equation 11.15 of the script) between a pair of sequences a_i, a_j

$$JC(a_i, a_j) = -\frac{3}{4} \ln \left(1 - \frac{4}{3} Ham(a_i, a_j) \right)$$

from the *probability-of-change* formula (equation 11.14 of the script) for the probability of an *observable* change occurring at any given site in time t

$$\mathbb{P}(\text{change} \mid t) = \frac{3}{4} \left(1 - e^{-\frac{4}{3}ut} \right).$$

2 Computing Hamming distances (1 point)

Write a program that reads in a set of aligned sequences and produces as output the corresponding Hamming distances matrix. To do so, modify the class `albi.phylo.Characters`.

3 Computing Jukes-Cantor distances (1 point)

Additionally, implement the Jukes-Cantor transformation of distances in class `albi.phylo.Characters`.

4 Jukes-Cantor simulator (3 points)

Write a program that takes as input a phylogenetic tree, a sequence length and a mutation rate and produces as output a set of simulated sequences. To do so, please modify the class `albi.phylo.JCSimulator`.

5 Run Neighbor-Joining on simulated data (3 points)

For each of the provided input files, choose three different reasonable values for L and u , use your simulator to generate sequences from them. Then, run the Neighbor-Joining program on the Hamming and Jukes-Cantor distances obtained from the simulated sequences. Compare the input trees with the output trees. How does the performance depend on the sequence length and mutation rate?

Due by 10am, Monday, 5 Feb 2007