Genome evolution – Ex 5 Due: June 20

Solve and submit this assignment in pairs, unless you have an ok from me.

In this assignment you are asked to implement the EM algorithm for simple trees and apply it to 3 genomes from the UCSC genome browser.

a) Select the genome triplet to work with. Go to http://genome.ucsc.edu/ and select any three species from the large vertebrate multiple alignment (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz17way/) or any three species from the fly multiple alignment. Use just one chromosome for your analysis.

b) Code something that can parse the (very simple) .maf files that contain the multiple alignment, if you want, you may write a script/program that will filter our all the species you are not using.

c) Implement the up-down algorithm

d) Implement the EM algorithm

e) Ignore gaps (filter away any loci that have a gap in one of the species)

f) Apply your code to 16 different collection of loci, defined by their flanking nucleotides in your "main" species (so set 1 would be loci that have an A before and A after, set 2 would be loci with an A before and C after, and so on).

g) Generate a report showing the different substitution probabilities you inferred. Try identifying and discussing possible differences between the matrices.

h) Submit your code, the results and a maximum 2 page summary of the findings in a tar.gz file to my email.

Good luck!