Genome evolution – Ex 3 Due: June 10

Work independently, or in pairs. You can use any available resource (books, papers, the web), but please cite what you use.

1. We showed in class how to solve the inference problem for simple trees given observations on the tree leafs. Assume the data for one of the species is missing at a certain position. In order to compute the total probability, can we run the inference algorithm as is, just replacing the up probability of the missing species with a uniform prior (Pr(x=A) = 0.25, Pr(x=C) = 0.25,..) ? Explain.

2. You are analyzing a set of aligned genomes of HIV strains. To escape recognition by the immune system, HIVs are hyper mutable in part of their genomes. You assume that the loci under study are divided into two subsets (which you don't know in advance). Each subset of loci is evolving under a different evolutionary regime (defined by substitution probabilities on each of the lineages).

a. Suggest a probabilistic representation to the problem (clearly you need more parameter than available in a simple tree model). Describe how to solve the inference problem in your model.

b. Develop an EM update formula (no need to prove it maximizes the Q function).

c. Assuming that adjacent loci are likely to evolve under the same regime; can you incorporate this into the model as well? How would you solve the inference problem in your combined model? (hint – it may be easier than solving a PhyloHMM model, since you may eliminate all cycles in your model)

3. The G+C content of a genomic region is defined as the fraction of Gs and Cs in it. You are observing a segment of 200 basepairs in two species. You assume a simple tree model that connects a single common ancestor species to the two observed species. The substitution matrices are symmetric, setting 1% substitution probability for any possible mutation at each of the lineages. You want to check if the G+C content of the region behaves as expected by the simple tree model. Specifically, if you assume the G+C content in species 1 is gc1 and the simple tree models is in effect, what should be the probability distribution for the G+C content on the other species?

b) (bonus) Now assume a model that includes two distinct regimes. The first regime works when the GC content is above 45% and is identical to what we described in (a). The second regime works when the G+C content is below 45% and is characterized by increased rate of mutation (2%) from G to A, while all other mutations occur at the same rate. In other words, according to this model, once crossing the 45% barrier, the mutational process prefers A's over G's. What will be the distribution of gc content in species 2 given species 1? (It is ok to solve this using simulation)

Good luck!