Genome evolution – Ex 2 Due: March 30

Solve 6 questions for 120 points. 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 transform an HMM model to a tree-like structure and how to perform inference efficiently using this representation. We also introduced n-order hidden markov models, in which the state at time t depends on the states at time t-1,t-2,..,t-n. Develop a polynomial algorithm for inference of single variable posterior probabilities in a 2-order hidden Markov model, or show the problem is NP-hard.(to remind we, we showed that just translating the 2-HMM to a Bayesian network create a non-tree structure). Bonus: can you generalize your conclusions to a polynomial algorithm for graphs with a certain property – state the claim without proving it.

2+3. In this question we assume that a sequence is generated from a 2-order Markov model and that we are trying to model it using a maximum likelihood 1-order Markov model. Assume throughout that you collected unlimited amount of sequence generated by the unknown 2-order model (hint – this means that you are only affected by the transition probabilities and stationary distribution of the model, and do not need an initial nucleotide or dinucleotide distribution).

a) Given a 1-order Markov model on the 4 nucleotides, what is the probability of observing a given 3-mer c_1 c_2 c_3? Compute the same for a 2-order Markov model.

b) Construct an example including a 2-order model θ^2 and a 3-mer c_1 c_2 c_3. Given a large sample from the 2-order model show what will be its maximum likelihood 1-order model $\theta^{1(ML)}$. Select model parameters and the 3-mer so that the ratio:

$$\Pr(c_1\ c_2\ c_3 \mid \theta^{1(ML)}) / \Pr(c_1\ c_2\ c_3 \mid \theta^2)$$

is as low as you can find (show you found the lowest ratio 3-mer for the specific model you have chosen). Remember that we are assuming you sample infinite amount of examples, so there are no questions of sample size or sample error.

c) Would the ML 1-order model for a data set necessarily maximize the minimal ratio in (b)? (Prove or give and example). (Bonus) Can you find an upper bound to the maximal ratio in (b)?

4. Prove explicitly that the EM for simple tress is monotonically improving the likelihood.

5-6. Modeling question: 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. Discuss inference in your model.

b. Develop an EM update formula (no need to prove it).

c. Assuming that adjacent loci are likely to evolve under the same regime; can you incorporate this into the model as well?

Good luck!