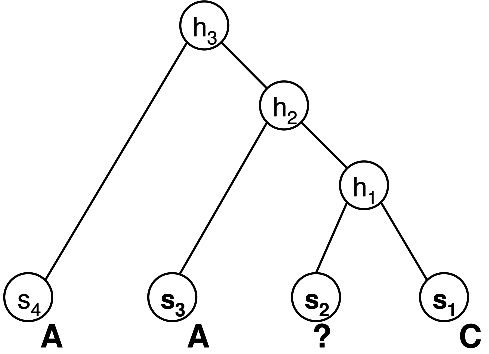
Genome evolution – Exam 1

Solve question 4 and 2 out of the first 3 questions. The C parts are a bit more difficult and you can get a good course grade without solving them. You should not use any reference material. Work independently.

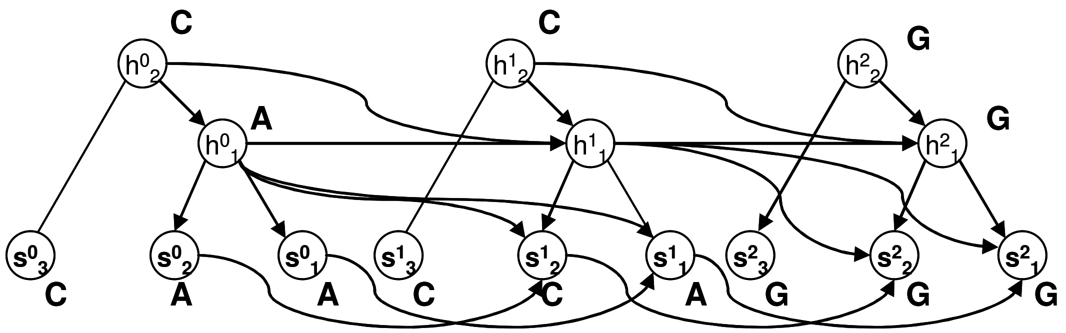
Question 1 (35 pt):
A) (20pt) Define the up-down inference algorithm for computing the likelihood of a multiple alignment given a tree model and observed sequence for a set of extant species. Prove the algorithm compute the likelihood correctly.
B) (10pt) Assume a simple tree model in which all conditional probabilities on all edges are set to 0.01 for mutations and 0.97 for conservation. Compute the posterior probability distribution on the hidden node h_2 given the model in figure 1 in which no data was available for species s_2. Develop a formula for $P(s_1, s_3, s_4 \mid \theta)$.
C) (5pt) Assuming you are given the phylogenetic model in figure 1, and an alignment that include only data on species s_1,s_3 and s_4. Write down a tree model on three species that would be completely analogous to the 4-species model: $P(s_1, s_3, s_4 \mid \theta^{new}) = P(s_1, s_3, s_4 \mid \theta)$. Qualitatively, to determine h_2 posterior, would you rather give up on data for species s_3 or species s_2



Question 2 (35pt):
A) (20pt) Given a Bayesian network P(x), define the general Gibbs sampling algorithm and prove it is giving rise to a reversible Markov Chain whose stationary distribution is P(x).
B) (10pt) How would your Gibbs algorithm sample a new value for a node h_1^1 in the PhyloHMM model shown in Figure 2, given the indicated current state, conditional probabilities that are all defined by:

$$\Pr(x_i^j \mid x_i^{j-1}, pax_i^{j-1}, pax_i^j) = \begin{cases} \mu & x_i^{j-1} = pax_i^{j-1}, x_i^j \neq pax_i^j \\ 2\mu & x_i^{j-1} \neq pax_i^{j-1}, x_i^j \neq pax_i^j \\ 1-3\mu & x_i^{j-1} = pax_i^{j-1}, x_i^j = pax_i^j \\ 1-6\mu & x_i^{j-1} \neq pax_i^{j-1}, x_i^j = pax_i^j \end{cases}$$

and μ=0.01.



C) (5 pt) You were led to believe mutations are more likely to occur in adjacent pairs. Assume that you are modeling an alignment of three species using the PhyloHMM model in figure 2, such that the strength of interaction between adjacent loci is determined by a uniformly distributed random variable $z \in [1..5]$. The joint distribution of the model is defined by:

$$P(s,h,z \mid \mu) = P(z)\prod_j \prod_i \Pr(x_i^j \mid pax_i^j, x_i^{j-1} pax_i^{j-1}, z)$$

Where:

$$\Pr(x_i^j \mid x_i^{j-1}, pax_i^{j-1}, pax_i^j, z) = \begin{cases} \mu & x_i^{j-1} = pax_i^{j-1}, x_i^j \neq pax_i^j \\ z\mu & x_i^{j-1} \neq pax_i^{j-1}, x_i^j \neq pax_i^j \\ 1-3\mu & x_i^{j-1} = pax_i^{j-1}, x_i^j = pax_i^j \\ 1-3z\mu & x_i^{j-1} \neq pax_i^{j-1}, x_i^j = pax_i^j \end{cases}$$

Write down the Gibbs update rule for the variable z given all other variables.

Question 3 (35pt):

A) (20pt): Define the general EM algorithm. Prove it is monotonically improving the likelihood.

B) (10pt): You are implementing an EM algorithm for motif finding which is similar to MEME, but allows only palindrome motifs (a palindrome is a motif that is reverse-symmetric – if the motif is of length l then m[i] equals the reverse complement of m[l-i]. Your model assume there is one hit (at position l_i) in each sequence, which is emitted from a PWM of length l. Write down the EM update rule (the solution to the maximization part of the EM) for this problem.

C) (5 pt): prove that your solution in B is correct.

Question 4 (30pt):

Codon bias: we learned that protein coding genes are structured in nucleotide triplets, which encode a polypeptide through a degenerate code (i.e. 4 or 2 codon can encode the same amino acid). This is a half-open question in which you are asked to apply one of the tools we developed in class to model this situation.

A) (20pt) Develop a probabilistic model for the evolution of a protein coding gene given different preferences for specific codons. Whatever your model is, you answer should be divided into definitions of (i) random variables (observed and hidden) (ii) model parameters (iii) joint probability (iv) inference technique to be used..

B) (5pt) Suggest an extension to your model in (a) when the mutation at each locus depends on the previous locus (As in phyloHMM) in addition to the codon bias, again build you answer as possible extension to parts (i)-(iv) from A.

C) (5pt) Suggest a strategy for checking if codon bias is under selection using data on derived allele frequencies for different degenerate codon positions and the stationary frequency of codons in the genome.