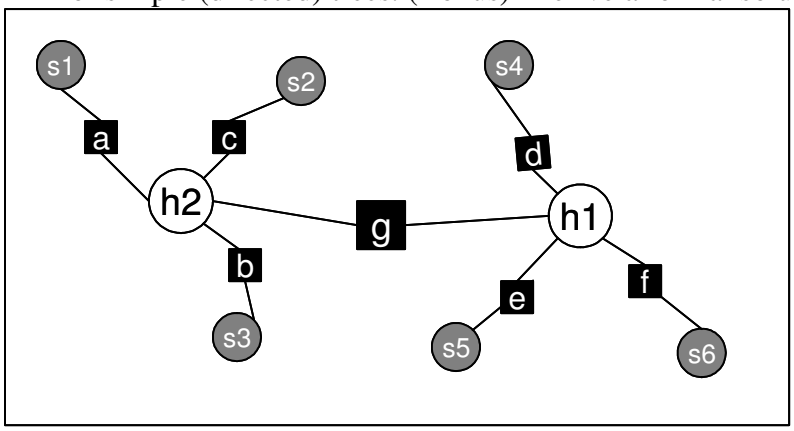
Genome evolution – Ex 3 Due: April 22

Solve 5 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. Develop a sampling based inference algorithm for a generic factor graph defined by the formula $P(x)=(1/Z)\Pi\phi(x_a)$. Your goal is to compute $P(x_i)$ for a single variable. Remember that in class we discussed algorithms for sampling from BNs, but that some of them were argued to work for undirected models as well.

2. We mentioned in class that the non-trivial partition functions of factor graphs make parameter estimation difficult (since the partition function depends on all parameters). Assume you are trying to implement an EM algorithm with an undirected model including 2 connected stars, using data on the star leafs s1,..,s6 (see drawing). You perform inference exactly, by enumerating over the hidden states of the two central nodes h1, h2, computing the potentials $P(x)=\Pi\phi(x_a)$ and the partition function Z. Write down the EM maximization formula $Q(\theta|\theta^k)$ assuming all factors a,b,c,d,e,f have the same parameterization and that you observed n(c1,c2,c3,c4,c5,c6) instances for each value combination of the observed variables s1,..,s6. Write down the extremum conditions for that function. Explain why solving these is not trivial as the EM for simple (directed) trees. (Bonus) -Derive a formal solution.



3. Find an example for a factor graph model that does not include zeros on any factor, but have two alternating LBP states that are not zero on any belief value. What is the mean field solution for your model?

4. Prove that the Bethe region-based entropy is exact for a model which is the uniform distribution. In other words, for any set of regions that are valid (have multiplier that sum up to 1 as explained in class), the region based entropy of a model that define the uniform distribution over all variables (H_R) is exactly the entropy for such a model ($H=\Sigma\ p(x)\log p(x)$) (read first sections from Yedidia et al. if you find this difficult).

5. a) Modeling question: the G+C content of a genomic region is defined as the fraction of G and C in it. You are observing a segment of 200 basepairs in two species. You assume a simple tree model that connect a single common ancestor species to the two observed species and indicate symmetric 1% mutation 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 before. 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, one crossing the 45% barrier, the mutational process prefers A's over G's. What will be the distribution P(gc2|gc1)? (It is ok to solve this using simulation)

Good luck!