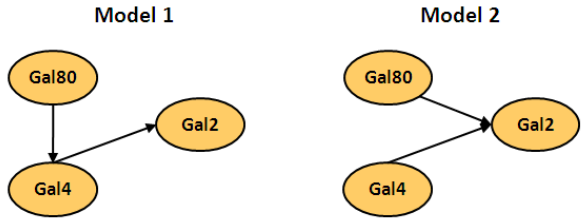
GENOME 541, Spring 2014
Problem Set #10

(Due June 12th 11:59am)

---

1. **[60 points] Model selection to find the best regulatory network**

   In this question, we will implement an algorithm for selecting among various structures of the regulatory network. Specifically, we will focus on two possible models of the galactose regulatory network in *S. cerevisiae*.



   Let's assume that expression levels are binary values (high, low), and we use table CPDs for both networks in Model 1 and 2.

   (a) **[10 points]** Describe what each parameter means.

   (b) **[10 points]** Say that we are given the gene expression data D measuring binary expression levels of the 3 genes (Gal80, Gal4 and Gal2) across 112 samples. Write down the likelihood function $L(\theta : D)$ for Model 1 and 2.

   (c) **[10 points]** Describe how to compute the maximum likelihood estimation of the parameters in Model 1 and 2.

   (d) **[25 points]** Download the data from `http://homes.cs.washington.edu/~suinlee/genome541/data/disc-gal80-gal4-gal2.txt`, and implement the code that computes the likelihood score for Model 1 and Model 2. Please submit the code and the resulting scores of Model 1 and 2.

   (e) **[5 points]** Select between model 1 and 2 based on the results in part (d).

2. **[40 points] Linear Module Network**

   Assume that you want to reconstruct the regulatory network of N genes from gene expression data. This data set contains expression levels of N genes in M experimental conditions. Denote by $x_i[1], ..., x_i[M]$, continuous-valued expression levels of gene i in M experimental conditions. You decided to use the Module network with K modules to represent the regulatory network. L genes were chosen to be candidate regulators based on the prior knowledge. Our goal is to learn the structure and the corresponding parameters of the Module network.

   (a) **[10 points]** Let's say that you use the linear model to express the dependency between x_i and its parents (linear CPD). Express the distribution of the expression level x_i based on its parents. Indicate what the parameters are.

(b) **[15 points]** Describe the algorithm for learning the structure of the model. Explain how to deal with a large number of candidate regulators and to select a small number of regulators with non-zero weights.

(c) **[15 points]** Given a fixed assignment of genes into K modules, write down the likelihood function of the model as a function of the parameters.