# I529: Bioinformatics in Molecular Biology and Genetics: Practical Applications **(3 CR)**

HW5 (Due: **April 29**, Friday, 5pm)

http://darwin.informatics.indiana.edu/col/courses/I529-16

## INTRODUCTION:

This homework consists of problems related to computational methods and algorithms. MS Word (doc) or Acrobat (pdf) are strongly encouraged for submitting answers. These files can also be submitted through Oncourse.

## QUESTION:

Don't hesitate to contact me (Haixu Tang: hatang@indiana.edu).

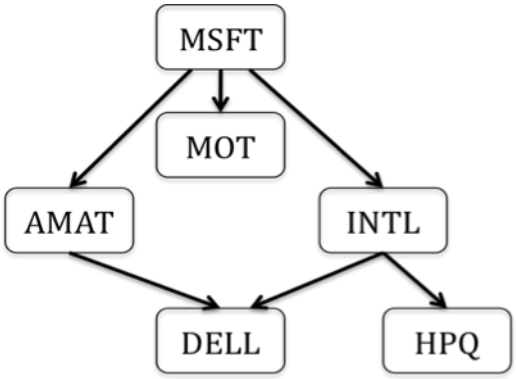
## INSTRUCTION:

1. Please start to work on the homework as soon as possible. For some of you without enough computational background may need much more time than others.
2. **Please ENJOY learning and practicing new things.**

You are NOT required to write programs. **100 points**

1. (20 pts) Assume an HMM θ is given. Devise a Gibbs sampling algorithm to find the observation symbol sequence x with the maximal total emission probability P(x|θ).

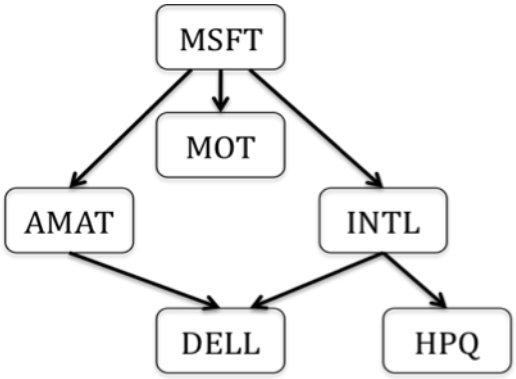
2. (25 pts) Consider the following Bayesian network.




We define three modules in the network: {MSFT}, {MOT, AMAT, INTL} and {DELL, HPQ}. 1) Devise the module network and corresponding conditional probability distributions on the module network. 2) Does the module network represent the same dependence structure as the original Bayesian Network? Why? 3) Is it true that any module assignment defined on this Bayesian network corresponds to a directed acyclic graph on the modules? If yes, explain why. Otherwise, give a counterexample.

3. (20 pts) The stock market data is actually over a time course, and it is most useful if we can predict the stock price from the previous marketing days. Devise a dynamic Bayesian network to achieve this on the six stocks shown above, assuming the price of each stock is dependent on the stock itself on the previous day and the stocks in the dependence structure as in the Bayesian network (but depending on the price of the stocks of the previous day). Briefly describe how the parameters in this model can be learned from real data.

4. Show the moralized graph of the following DAG (15 pts).



5. (20 pts) Consider the Bayes net shown below. Here, the nodes represent the following variables: $X_1 \in$ {winter, spring, summer, autumn}, $X_2 \in$ {salmon, sea bass}, $X_3 \in$ {light, medium, dark}, $X_4 \in$ {wide, thin}. The condition probability tables are, $p(X_1)=(0.25,0.25,0.25,0.25)$, $p(X_2|X_1)=(0.9, 0.1; 0.3, 0.7; 0.4, 0.6; 0.8, 0.2)$, $p(X_3|X_2)=(0.33, 0.33, 0.34; 0.8, 0.1, 0.1)$, $p(X_4| X_2)=(0.4, 0.6; 0.95, 0.05)$. Note that in $P(X_2|X_1)$, the values between ";" represent the probability distributions under different conditions of $X_1$, so that they should summed up to 1. Thus, $P(X_2=\text{salmon}|X_1=\text{winter})=0.9$, and $P(X_2=\text{sea bass}|X_1=\text{summer})=0.6$.

(1) Suppose the fish was caught on December 20 – the end of autumn and the beginning of winter – and the $p(X_1)=(0.5,0,0,0.5)$, instead of the above prior. Suppose the lightness has not been measured but it is known to be thin. Classify the fish to be salmon or sea bass.

(2) Suppose we all know is that the fish is thin and medium lightness. What season is it now, most likely? Use the prior $p(X_1)=(0.25,0.25,0.25,0.25)$.