Problem 1:

Download the cancer vs. normal gene expression data set for lung cancer (see course web page). This data set contains 15 tumor samples and 8 normal samples (one sample per column). Rows correspond to genes.

A. Implement a suitable statistical test to identify genes that are overexpressed in tumors as compared to normal samples (attach your code to your submission, can be in any programming language or matlab or R). What are the top 25 genes you found? Is there any way you can do a "sanity check" to see that you found genes that are actually differentially expressed (describe how you would do it)? (submit: list of genes, your code, and description a "sanity check" test)

B. Describe (don't implement) a quantitative evaluation scheme for your statistical test (either on biological or synthetic data or in some other way - up to you). Do not go into minute details, tell us what your gold standard is in this evaluation, why you chose to evaluate this way, what type of results you'd report. Does your evaluation have any drawbacks - what are they? (submit: description of your evaluation scheme and discussion of its advantages and drawbacks)

C. Does your statistical test in (A) depend on any assumptions? What are they? Discuss how you could design a different test that does not make those assumptions. What are the advantages and drawbacks of your new test? Are there situation in which your test in (A) is better than the one in (C)? If yes, which situations and why.

Some possible references (you don't have to read all of these or implement algorithms that are as complicated as the ones suggested in these papers, but if you are lost, these may give you some ideas):

Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res. 2001 Jul;11(7):1227-36.

Ideker T, Thorsson V, Siegel AF, Hood LE. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J Comput Biol. 2000;7(6):805-17.

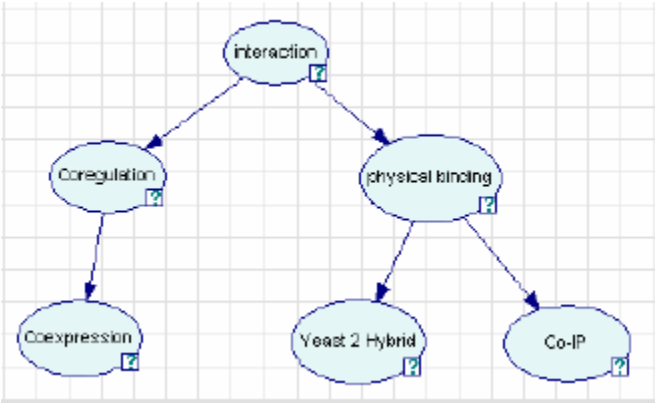
Park T, Yi SG, Lee S, Lee SY, Yoo DH, Ahn JI, Lee YS. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinformatics. 2003 Apr 12;19(6):694-703.

Problem 2.

You are interested in predicting protein-protein interactions in baker's yeast based on yeast two hybrid, co-immunoprecipitation, and gene expression microarray data. You decide to use Bayesian networks to combine these methods.

1. You need a structure and parameters for the network. Luckily, in your courses in graduate school you've learned that Bayesian networks can be either expert systems or can be learnt. With this particular problem in mind, discuss what are advantages and disadvantages in learning vs. constructing based on expert opinion: A) structure of the network and B) parameters of the network. (Hint: think of properties of learning, flexibility of the algorithms, pitfalls of algorithms)

2. Either through learning or by expert opinion (depending on your answer above), you end up with the following simple structure. This network considers a pair of proteins at a time, iterating through all possible combinations of two proteins that you want to consider, and determines the confidence level of these two proteins having an interaction based on the three types of data we are using. The data includes information on whether pairs of proteins from the genome have a positive yeast two hybrid experiment, whether they have a positive co-immunoprecipitation experiment, and a gene expression dataset with expression of all proteins in the genome over a large set of conditions (50-100 experiments. Nodes in the network are discrete, and yeast two hybrid and Co-IP nodes assume binary input.



A. Discuss details of how you would process gene expression data to get it into the form that the network can accept (i.e. what exactly will the input into the Coexpression node be, and how will this input be binned into discrete categories)?

B. You decided to first use the network as an expert system. Write out conditional probability tables of each node (parameter sets). You can consider yourself the expert who provides the numbers to put in the tables (you won't be graded on the correctness of the numbers, only on the correctness of understanding how CPTs are constructed).

C. Let's say you initialized the network with input values for two proteins. Now you want to calculate the confidence level for these two proteins having an interaction. Write out formulaically (don't use the numbers from B, just symbols) how you will calculate P(interaction) based on the data you have, a parameter set, and this network structure (remember to use the structure to take into account conditional independence relationships). If you are confused by Bayesian

networks, this link may be helpful:
http://www.ai.mit.edu/~murphyk/Bayes/bnintro.html

3. Now you decide to learn the parameters for the network above. What data could you use for learning the parameters (i.e. what are your "answers")? Discuss how you would set up learning and evaluate how your network is doing. Discuss advantages and pitfalls of your setup, and assumptions you are making, if any.

Problem 3.

Pick a paper in the area of genomic data integration, microarray data analysis, comparative genomics, prediction of pathways and networks in genomics, or visual analysis of genomic data. The paper you pick has to be no older than 2002 (publication date). It may have significant biological content, it may have a light biological content, but it has to have a significant computational content (use papers in class as a guide, if in doubt, e-mail me the abstract and I will let you know). Write a critique of the paper – no more than 1.5 pages, single spaced, 12pt font (less is fine), discuss what problem the authors addressed, what they accomplished, how, how did they evaluate their method. Make sure to discuss advantages/drawbacks of their method and suggest some future work (beyond what they talk about in the paper).

If you are at a loss picking a paper on an appropriate topic, try journals: Bioinformatics, BMC Bioinformatics, Nature Genetics, Science, Genome Research (these aren't the only ones you can use, these are just suggestions).