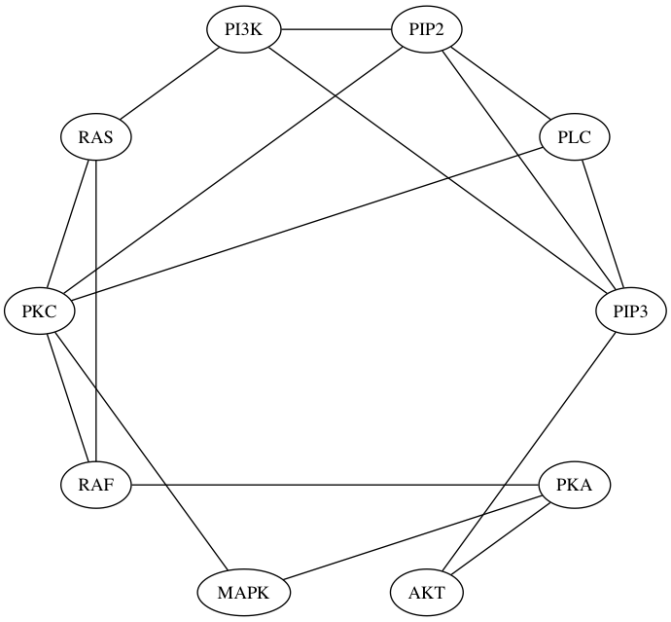
BIOE 298, SECTIONS MFI & B

HOMEWORK 6

Due 4/26/2018 by end of class.
Upload a single PDF with your answers to both parts to Compass.

Part I: Machine Problem (50 points)



The above figure depicts protein-protein interactions in a human signal transduction network. Your goal is to find the most central and least central proteins in this network.

(1) Construct an adjacency matrix to represent the connections in the network.
   - An adjacency matrix is a square matrix **A** such that each element a_{ij} equals 1 if node i is directly connected to node j.
   - Since the above graph is undirected, your adjacency matrix should be symmetric. If $a_{ij} = 1$, the $a_{ji} = 1$.
   - The diagonal elements (a_{ii}) must be left zero, since no node is "connected" to itself.

(2) Calculate the leading eigenvector for the adjacency matrix. The leading eigenvector is associated with the eigenvalue with the largest magnitude.

(3) Using the magnitude of the entries in the leading eigenvector, report the most central and least central proteins in the network. How does the centrality of these proteins compared with the number of connections involving these proteins? Is the most central protein always the protein with the largest number of direct connections?

Part II: Machine Problem (50 points)

A team of researchers used DNA microarrays to measure gene expression in a large set of breast cancer cell lines (Kao, et. al, PLoS ONE 4(7): e6146. doi:10.1371/ journal.pone.0006146). In this exercise, you will use gene expression profiles from this study and Principal Components Analysis to identify genes whose expression distinguishes invasive (IDC) from regular (DC) ductal carcinoma.

(1) Load the mat file `HW6_data.mat`. Note that the data in Homework 6 have been slightly modified from the previous assignment (Homework 5). Be sure to download the Homework 6 file, which contains the following variables:
   - `training_lines` is a Matlab table containing gene expression data for the IDC and DC cell lines. Each of the 8750 rows corresponds to a gene with variable expression across the cell lines. Each of the **21** columns represents a cell line.
   - `is_invasive` is a 21-element vector signifying if the columns in `training_lines` represent invasive (1) or regular (0) ductal carcinoma.
   - `gene_names` is an array of names for the genes (rows) in the table `training_lines`.

(2) Use the `pca` function to calculate principal component loadings, scores, and explained variances for the data. Note that the `pca` function may only return the first 20-30 principal components (rather than all 8750).

(3) Plot the data for the cell lines along the first two principal components. (Use the `scatter` function with four arguments to color the points as either IDC or DC.)
   - How much of the total variation is shown on this plot?
   - Do either (or both) of the principal components separate the IFC and DC samples?

(4) If any of the first two principal components separate the IFC and DC samples, find the gene whose expression is increased the most in IFC vs. DC, and the gene whose expression is decreased most significantly between IFC and DC.

Remember to submit all code, outputs, and explanations for these problems.