Research Summary of Graduate Student of Hoque-Lab: Chris Kives

Short project: **GRNN (Generalized Regression based Neural Network) and kNN based Hybrid Approach for Missing value Estimations in DNA Microarray**

Gene expression is biological data that is collected through a physical device called a DNA microarray, which can benefit from the quick computation provided by computers for various reasons. It is studied, so researchers can identify which genes are ”turned on” or ”turned off,” which means which genes are actively producing mRNA. The information gathered about gene expression data allows researchers to make a comparison between two different cells functionality, because it shows the variation between the quantities to which each gene is expressed between the cells.

DNA microarrays can examine thousands of genes under many different conditions, simultaneously, with one physical chip. The process involves extracting mRNA from cells. Then complimentary DNA (cDNA) is constructed from the mRNA, and mixed with a florescent molecule that will bind to it. The chip contains microscopic spots with many single strands of DNA per spot and with one unique gene per spot. The cDNA is then added to the microarray, and it starts binding to the corresponding single stranded DNA gene at each spot. The florescent level at each spot is then scanned by a computer, which then calculates the expression level at each spot. This entire process basically allows researchers examine the mRNA expression levels for tens of thousands of genes simultaneously. The data read in by the computer from the microarray can be then constructed into a matrix of values. For the methods described below, the genes are row vectors and the samples are column vectors of the matrices.

Motivation: The entire process of reading DNA microarrays is not completely without any errors, and to make conclusions meaningful and consistent, data often has to be complete. Microarrays are costly to produce; therefore, methods must be used to keep the results of the microarray without having to redo the entire experiment.

Clearly, something must be done about the missing values. Initially, the missing value is flagged as needing correction. But how would we correct it?

One method is to throw out the data completely, but this means an entire gene or sample becomes useless for analysis. This is clearly not desired.

Another approach is to zero out the missing values. This approach allows us to use the values, but any comparisons involving those values become meaningless. This is also not desired.

A third approach could be averaging the data of the other samples’ gene to obtain the missing values, which provides a little better results. However, this still isn’t too useful, because the average of the genes per sample isn’t a good metric for the current sample most likely. So this method isn’t too effective.

What is needed is a robust method to accurately predict missing values, so missing value estimation techniques must be used. Missing value estimation is a method that attempts to find an approximate value for missing values in data sets, which in our case is a microarray data stored in a matrix. Luckily, microarray data many gene expression levels are related, so with available information educated guesses on the correct value for the missing values.

The methods can either use local, global, or external knowledge for finding the solution. Local approach grabs a small sample of similar genes to compare to the one with missing values. Global approach uses available data from the entire matrix to arrive at a solution. Finally, external knowledge approach uses data outside the matrix to determine the missing values. What follows is an examination into at least one method in each of these approaches.